KJCIS

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

ISSN: 2616-9592

Volume: 2 | Issue : 2 | July-December | 2019

KIET JOURNAL OF COMPUTING AND INFORMATION SCIENCES

Volume 2, Issue 2, 2019
ISSN: 2616-9592
Frequency Bi-Annual

College of Computing & Information Sciences
Karachi Institute of Economics & Technology

College of Computing & Information Sciences

Vision

To develop technology entrepreneurs & leaders for national & international market

Mission

To produce quality professionals by using diverse learning methodologies, aspiring faculty, innovative curriculum and cutting edge research, in the field of computing & information sciences.

AIMS AND SCOPE

KIET Journal of Computing and Information Sciences (KJCIS) is the bi-annual, multi-disciplinary research journal published by **College of Computing & Information Sciences (CoCIS)** at **Karachi Institute of Economics and Technology (KIET)**, Karachi, Pakistan. **KJCIS** aims to provide a panoramic view of the state of the art development in the field of computing and information sciences at global level.

It provides a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and information sciences to share their findings and contribute to the knowledge domain at global level. The journal also fills the gap between academician and industrial research community.

KJCIS focused areas for publication includes; but not limited to:

- Data mining
- Big data
- Machine learning
- Artificial intelligence
- Mobile applications
- Computer networks
- Cryptography and information security
- Mobile and wireless communication
- Adhoc and body area networks
- Software engineering
- Speech and pattern recognition
- Evolutionary computation
- Semantic web and its application
- Data base technologies and its applications
- Internet of things (IoT)
- Computer vision
- Distributed computing
- Grid and cloud computing

OPEN ACCESS POLICY

For the benefit of authors and research community, this journal adopts open access policy, which means that the authors can self-archive their published articles on their own website or their institutional repositories. The readers can download or reuse any article free of charge for research, further study or any other non profitable academic activity.

PEER REVIEW POLICY

Peer review is the process to uphold the quality and validity of the published articles. KJCIS uses double-blind peer review policy to ensure only high-quality publications are selected for the journal. Papers are referred to at least two experts as suggested by the editorial board. All publication decisions are made by the journal's Editors-in-Chief on the basis of the referees' reports. We expect our Board of Reviewing Editors and reviewers to treat manuscripts as confidential material. The identities of authors and reviewers remain confidential throughout the process.

COPYRIGHT

DISCLAIMER

The opinions expressed in **KIET Journal of Computing and Information Sciences (KJCIS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board and the editorial board. Papers published in KJCIS are processed through double blind peer-review by subject specialists and language experts. Neither the **CoCIS** nor the editors of **KJCIS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead; errors should be reported directly to the corresponding authors of the articles.

Editorial Board

Language Editors

Production Management

Table of Content

Identity Lock – Privacy-Preserving Data Publishing Tool

Neha Maroof Siddiqui[1] Sara Benish[2] Tehreem Qamar[3]

Abstract

In today's world, data sharing is very common. Currently, the strong movement is occurring towards publishing data for statistical studies. In this case, data publishers are providing some sort of data to the research field, but they do not know what kind of things the 3rd party can do with the data provided to them. Data preservation is an important aspect when sharing data because attackers can easily disclose a person's identity and their personal information. Hence, in order to secure privacy, different methodologies are implemented on data. This paper presents Identity Lock - the Privacy Preserving Data Publishing (PPDP) tool, uses various anonymization techniques and implements k-Incognito, l-Incognito and ε-Differential Privacy algorithm to hide and anonymize data. The software also performs the experimental evaluation in order to calculate the performance of the algorithm on the basis of how much utility and privacy is maintained.

Keyword: Privacy Preserving, PPDP, Data Preservation, Anonymization, Privacy Models

1 Introduction

Demand of microdata is becoming diverse. Organizations collect and share this microdata for knowledge-based decision making [1]. In statistics, microdata is a set of records containing personal information [2]. It contains some personal information that causes privacy issues. Not only that, study shows that 87% of United States population was identified easily from published datasets [3].

Statistical studies like enumeration, population factors, health statistics and road accidents records, all created from data. While publishing data, privacy concern and preservation is considered as an imperative factor for viable use of data. This information is kept in electronic configuration, without causing any trouble to a person [3]. The consistently increasing velocity of data makes security a challenging task, particularly when the data is high in storage.

In 2006, Netflix an online DVD-rental company released their data to improve its movie recommendation algorithm [4]. The company released anonymized data, but just 16 days later two specialists from The University of Texas easily distinguished clients by coordinating the informational data from other sites like IMDb. On December 17, 2009, four Netflix clients documented a legal claim against Netflix, asserting that Netflix had disregarded U.S. reasonable exchange laws and the Video Privacy Protection Act by releasing the datasets [5]. Another case of sensitive information leakage occurred when AOL (American Online)- an online service provider released their search log of 657,000 American citizens from which an individual

[1] Jinnah University for Women, Karachi | nehamaroofsiddiqui@gmail.com
[2] Jinnah University for Women, Karachi | sarabenish@yahoo.com
[3] Jinnah University for Women, Karachi | tehreem_qamar10@hotmail.com

named Thelma Arnold was identified. Later searched data was removed from the websites but the damage is already done [6].

Generally, privacy concerns are identified with validation, data accessing, data encryption, and data publishing. Numerous data holders distribute the microdata of their organization for various purposes such that released data don't violate a person's privacy [4]. To address these serious privacy violations, data should be published after certain anonymization processes are applied to it. The research area focusing this issue is known as PPDP (Privacy-Preserving Data Publishing). It is an important step for securely publishing microdata for research analysis and statistical studies. Until now, different methods [5],[6],[7] are proposed that mainly focused on protecting the disclosure of private information, while providing the utility in published data.

The contribution of this paper can be summarized as:

- Survey of different anonymization algorithm and their limitations in preserving privacy and providing utility data.
- Development of a tool named "Identity Lock" that implement the different algorithm (i.e. k-Incognito, l- Incognito and ε-Differential Privacy)
- Evaluation of algorithms on a variety of dataset in order to keep the privacy of each individual and provide utility data for further statistical need.

The rest of the paper is organized as follows: Section II presented the background of PPDP extended by some general privacy techniques and survey of related work on different privacy models. The proposed system is presented in Section III with the experimental analysis in Section IV, while the conclusion discussed the final verdict of the research study in Section V.

2 Background & Related Work

A Privacy-Preserving Data Publishing

The approach of analyzing and acting upon data is extremely important for various organizations [8]. The sharing of data may lead to misuse or excessive data distortion. Privacy-Preserving Data Publishing is a concept providing method for publishing useful information while preserving individual's privacy. Figure 1 described the process of publishing privacy preserved data by following steps presented as:

- An owner collects raw data from their organizations.
- Anonymization techniques are applied to preserve data privacy.
- Once the privacy is preserved, data is released to publish for research and statistical analysis.

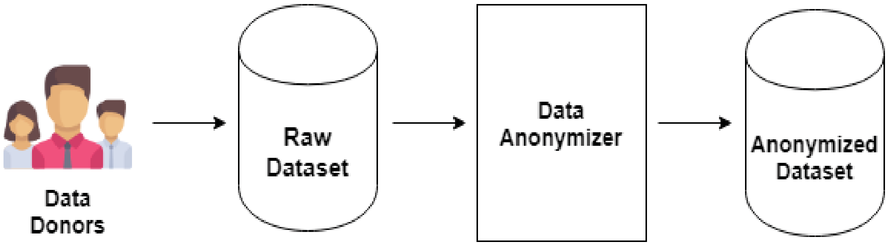


Figure 1: Privacy Preserving Data Publishing Architecture.

For the process of anonymization, microdata is generally categorized into three forms:

- Direct identifier: There are some attributes that easily identify a person's identity such as name, address, user ID, etc.
- Quasi Identifier (QI): The group of attributes which helps in recognizing a person such as class, age or gender.
- Sensitive Attribute: The data fields that contain an individual's personal information such as disease, salary.

B General Techniques

The most general techniques use to anonymize data are:

1) Generalization: It is a process that transforms the group of records into more generalized one. It removes direct identifiers from the datasets and then assigns a common value to the group of data records that possess the same kind of data [9]. It is one of the flexible technique [10] but it lacks in providing data utility [11].

2) Bucketization: It aims to preserve privacy by dividing records according to a quasi-identifier and assigns a unique ID to each division[12]. Then, both the quasi-identifier and sensitive value in records are published separately. By applying this process, specific values are not lost, but it breaks the relation of QI and sensitive attribute [13].

3) Suppression: Similar to generalization, suppression first removes the direct identifiers, then changes specific values of quasi-identifier (QI) to *, completely hides some values indicating that replaced record is not meant to be published [14]. The replacement of values with "*" causes information loss which is the main drawback of this method [15].

4) Perturbation: This technique is based on randomization [16]. It can be implemented by replacing the original value with any random value. The perturbed data records change sensitive values while quasi-identifier remains unchanged due to which resultant dataset does not ensure privacy protection [15].

5) Slicing: It is done on records by dividing datasets horizontally and vertically[17]. Vertical partitioning grouped co-relative attribute in a column and horizontal

partitioning grouped sets of records in buckets. Each bucket is then randomly permutated.

Table 1: General Techniques

Techniques	Advantages	Disadvantages
Generalization	It protects identity disclosure by replacing specific information.	It is prone to homogeneity attack and the background knowledge attack.
Bucketization	It prevents dataset from record linkage attack by assuming a clear separation in between QIs and SAs.	It failed to prevent membership disclosure i.e. doesn't protect attribute disclosure to sufficient extent.
Suppression	It provides identity disclosure risk by suppressing the real value.	"*" value disturbed utility at high rate, especially during statistical analysis.
Perturbation	The attacker cannot perform the sensitive linkages or recover sensitive information from the published data.	The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records.
Slicing	It preserves better utility than that of generalization while protecting privacy.	Due to the correlation of high attribute, privacy violation may happen in slicing technique.

C Related Work

K-anonymity was first proposed by Sweeney [18] in which she replaces values in the dataset by less-specific value. After which, the Datafly approach [19] uses a heuristic to perform generalization on quasi-attributes. However, again, no formal foundations or abstractions have been provided. Samarti's work [20] uses k-minimal generalization but failed to maintain optimal minimum information loss. The study in [20], [21] on the method is known as minimal generalization which is independent of the purpose of released data. Another approximation algorithm was proposed by El-Amawy[22] which provides optimal anonymization. However, the method failed when larger values of k are desired. Moreover, [6] presented a method which uses a clustering mechanism but results in minimizing the utility of data as suppression hide most of the information.

Fung et al. in his work [23] discussed the main goal of the data release by implementing classification, resulting in k-anonymize data that is optimal and minimizes the cost metric. Generally, achieving the optimal k-anonymity is NP-hard [22], [24].Besides the general anonymization techniques, LeFevre et al. Studies an extension of k-anonymization [25] and proposed the multidimensional k-anonymity. Moreover, LeFevre et al. [26] broadened the previously stated multidimensional approach for anonymizing a specific task i.e. classification. Xu et al. [27] discussed different greedy approaches to use k-anonymization for cell generalization. His work proved that the anonymized microdata results in less utility loss than that of used by LeFevre et al. [26].

Apart from k- anonymization, l-diversity [28] save both the data privacy and its utility. It obtains anonymization with the diversity of sensitive values on quasi-identifying groups. In Bucketization [29], publishing sensitive values separately from the quasi-identifier, secure linking attacks and maintain data utility as no changes are made on specific values. But still, the identity of the victim can easily be known by the attacker as it does not provide membership disclosure [30]. Further Li et al. [31], discusses the limitations of l-diversity and introduced a t-closeness technique to overcome privacy attacks. The t-closeness [32] calculates Earth Mover Distance (EMD) between two distributions for all the field contained in the dataset and a sensitive attribute. In t-closeness, there is a correlation between sensitive attributes of a dataset and QIDs; t-closeness degrades utility of privacy preserved data. Secondly, for sensitive numerical data, the t-closeness is unable to prevent attribute linkage attacks [33]. Thirdly, t-closeness uses EMD measurement that is not perfect and flexible enough to impose different privacy levels on different the sensitive attributes.

Perturbation used in [34], added noise to datasets. Xiao [35] and Chaytor [36] state that perturbation is only suitable in cases that focus on preserving privacy while false records don't affect research analysis result. Chaytor and Wang [36], work on randomly assigning sensitive attribute by dividing its domain. It results in high error where small ranges are used. Following the randomization and perturbation, Laplacian noise is included in differential privacy [37] to improve the sensitivity of data.

The general loss metric is presented in [38] uses a generalized data to calculates a normalized loss of information. Dewri et al. [39], use a weighted k-anonymity, focusing on the privacy-utility issue for better results.

Table 2: Summary of Privacy Models

Authors	Description
Sweeney [12]	K-anonymity, presented the protection model to anonymize data implementing generalization and replacing values in dataset by less-specific value. But the approach used lacks in securing individual's privacy.
Sweeney [17]	Data fly, applied heuristic approach to perform generalization on quasi-attributes. It guarantees k-anonymous transformation but doesn't provide the minimal generalization.
Samarti, Sweeney [18], [19]	Minimal generalization, implemented generalization and suppression on datasets but generalize data more than its needed and failed in providing optimal minimum information loss.
P. Kulasinghe [20]	The approximation algorithm, proposed to provide optimal anonymization. However, method failed when larger values of k are desired.
G. Aggarwal [21]	Clustering mechanism, maintain k-anonymity. However, this approach failed to maintain the utility of data, as suppression hide most of the information useful for data mining or research work.

Fung et al. [22]	Classification, presented a top-down approach to iteratively refine the data from a general state into a special state. However, the leakage of two data points can result in complete disclosure of information.
LeFevre et al. [24]	Mondrian Multi-dimensional anonymity, where generalization is performed on multi-dimensional data using approximation algorithm. But it lacks in providing as much data utility as needed for research studies
Xu et al. [26]	Greedy approximation algorithm, to use k-anonymization for cell generalization. His work proved that the anonymized micro data results in less utility loss.
Ercan Nergiz [27]	δ-presence, published sensitive values separately from quasi identifier, secure linking attacks and maintain data utility.
Li et al. [29],	T-closeness, discusses the disadvantages and limitations of l-diversity and then introduced t-closeness technique to overcome various privacy attacks. However, this technique does preserve feature disclosure but identity is still disclosed.
Xiao et al [33]	Optimal perturbation, achieves anonymization with a multi-level perturbation approach that release multiple data sets anonymized on different levels of privacy.
Chaytor and Wang [34]	Small domain generalization, work on randomly assigning sensitive attribute by dividing its domain. This approach retains more data yet not guaranteed information disclosure risk.

3 Identity Lock

Identity Lock is designed to implement k-Incognito [40], l-Incognito [41] and ε-Differential Privacy [42]. We have chosen these algorithms on the basis of the following facts:

- Chosen algorithms are extensively cited.
- These algorithms use different strategies that work on both categorical and numerical attributes.

It also evaluates the algorithm on the basis of utility and privacy maintained by the algorithms. Figure 2 describes the software design of Identity Lock. The software requests a file of supported format from a user. As soon as user uploads the file to be anonymized, the software reads data from the uploaded file, display it in GUI and ask a user to choose sensitive attribute (SA) and a privacy model to proceed further. Once the required parameters are selected, the software starts anonymizing the given dataset by following the privacy model selected by the user.

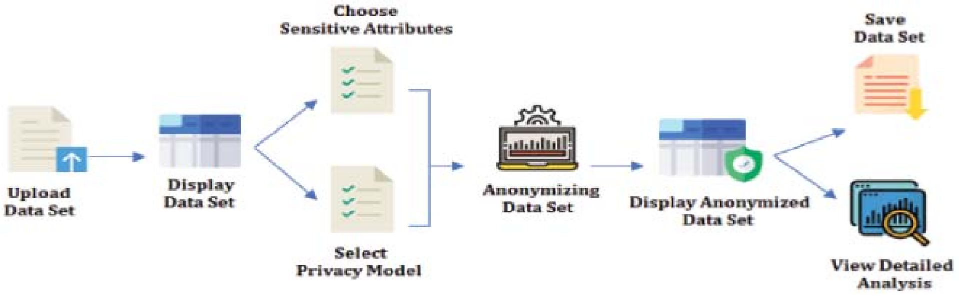


Figure 2: Processing Model of Identity Lock.

Identity Lock supports the following algorithms: k-Incognito, l-Incognito, and ε-Differential Privacy

- K-Incognito uses bottom-up search to anonymize the quasi attributes in datasets. It works on protecting identity disclosure. Yet the privacy is vulnerable when attacker have strong background knowledge of individual [40].

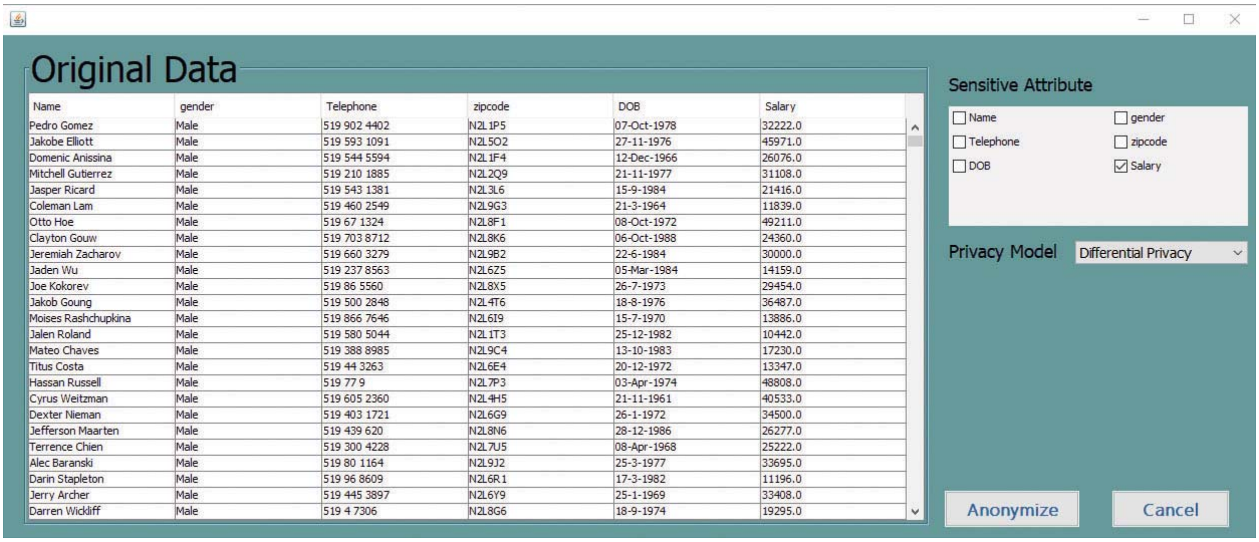
- To overcome such problems, l-Incognito diversify the values of sensitive attributes within an equivalence class. It protects datasets against attribute disclosure [41].

- For statistical data, ε-Differential Privacy uses additive noise approach to ensures privacy [42].

- Levenshtein metric is implemented to measure utility maintained by anonymized dataset [43].

- Shannon Entropy metric is used to measure the privacy of anonymized dataset [44].

- To compare execution time of an algorithm, the software monitored time from start till the end of the process.



Original Data

Name	gender	Telephone	zipcode	DOB	Salary
Pedro Gomez	Male	519 902 4402	N2L1P5	07-Oct-1978	32222.0
Jakobe Elliott	Male	519 593 1091	N2L5O2	27-11-1976	45971.0
Domenic Anissina	Male	519 544 5594	N2L1F4	12-Dec-1966	26076.0
Mitchell Gutierrez	Male	519 210 1885	N2L2Q9	21-11-1977	31108.0
Jasper Ricard	Male	519 543 1381	N2L3L6	15-9-1984	21416.0
Coleman Lam	Male	519 460 2549	N2L9G3	21-3-1964	11839.0
Otto Hoe	Male	519 67 1324	N2L8F1	08-Oct-1972	49211.0
Clayton Gouw	Male	519 703 8712	N2L8K6	06-Oct-1988	24360.0
Jeremiah Zacharov	Male	519 660 3279	N2L9B2	22-6-1984	30000.0
Jaden Wu	Male	519 237 8563	N2L6Z5	05-Mar-1984	14159.0
Joe Kokorev	Male	519 86 5560	N2L8X5	26-7-1973	29454.0
Jakob Goung	Male	519 500 2848	N2L4T6	18-8-1976	36487.0
Moises Rashchupkina	Male	519 866 7646	N2L6I9	15-7-1970	13886.0
Jalen Roland	Male	519 580 5044	N2L1T3	25-12-1982	10442.0
Mateo Chaves	Male	519 388 8985	N2L9C4	13-10-1983	17230.0
Titus Costa	Male	519 44 3263	N2L6E4	20-12-1972	13347.0
Hassan Russell	Male	519 77 9	N2L7P3	03-Apr-1974	48808.0
Cyrus Weitzman	Male	519 605 2360	N2L4H5	21-11-1961	40533.0
Dexter Nieman	Male	519 403 1721	N2L6G9	26-1-1972	34500.0
Jefferson Maarten	Male	519 439 620	N2L8N6	28-12-1986	26277.0
Terrence Chien	Male	519 300 4228	N2L7U5	08-Apr-1968	25222.0
Alec Baranski	Male	519 80 1164	N2L9J2	25-3-1977	33695.0
Darin Stapleton	Male	519 96 8609	N2L6R1	17-3-1982	11196.0
Jerry Archer	Male	519 445 3897	N2L6Y9	25-1-1969	33408.0
Darren Wickliff	Male	519 4 7306	N2L8G6	18-9-1974	19295.0

Sensitive Attribute

□ Name □ gender
□ Telephone □ zipcode
□ DOB ☑ Salary

Privacy Model: Differential Privacy

Figure 3: Uploaded Dataset

Identity Lock also featured the automatic anonymization of the dataset based on the privacy model which maintained good utility-privacy of data. This feature implements all the algorithms, compares utility and privacy maintained by each algorithm and then display result that best guarantees both the utility of data and privacy of the individual's in a dataset. The anonymized data can then be saved in .csv or .xls file format. The user can also view a detailed analysis of utility and privacy maintained by the privacy model and the time taken by the system to anonymize the whole dataset. The detailed mechanism of algorithms is discussed below:

A K-Incognito

K-Incognito technique follows the global-recording model called a full-domain generalization. DGH (Domain Generalization Hierarchy) [40] is formed for each QI attributes Q. The number of valid domain generalization[45] varies by the depth of its VGH (Value Generalization Hierarchy). For a dataset containing multiple QI attributes, the domain generalization hierarchies of each individual attribute are merged to build a generalization lattice.

Incognito algorithm[7] uses bottom-up search to pass over this generalization lattice. It starts by checking single-dimensional nodes of QI attributes, proceeds by iterating to an increasing subgroup of QI sets in the lattice to check k-anonymity requirement. If a node fulfils the property of k-anonymity, then all of its direct generalizations is marked, guaranteeing that they also satisfy the property of k-anonymity. The algorithm terminates when all the combinations of QI attributes have been considered.

This method may be in-efficient with respect to time but the anonymized dataset contains the maximum quantity of information that makes this algorithm an optimal solution[45] for preserving privacy.

B ε-Differential Privacy:

Differential privacy is considered as "State-of-Art" technique in the data privacy field. Noise addition technique was first outlined by Dwork [42] to address the anonymization of statistical data. It guarantees privacy by utilizing noise addition perturbation methods that transform sensitive attribute by adding calculative noise to it. The differential privacy method ensures that an attacker can't succeed in misusing information about any person in the dataset.

According to Dwork [46], if two datasets D1 and D2 differ or disagree in a single record, an anonymized algorithm A is said to satisfy ε-differential privacy, if it results R supports the equation:

$$\frac{P[A(D_1)\epsilon R]}{P[A(D_2)\epsilon R]} \leq e^{\epsilon} \qquad (1)$$

Where P represents the probability of an event occur and refers to the statistical distance to determine the strength of privacy. It has been noted by C.Dwork[42] that smaller values for ε give more privacy while ε= 0 is said to be completely differential private. However, utility risk increases with smaller ε value.

C L-Incognito

K-anonymity privacy definition is vulnerable to adversaries that have strong background knowledge of individuals represented in the dataset. [41] l-diversity tries to overcome such vulnerabilities by diversifying sensitive values within each equivalence class.

The toolbox implements an approach to multiple-sensitive attributes relies on the solution described by Sweeney [9]. All sensitive values should merge into a single attribute, which then specified as the only sensitive attribute. Incognito anonymization with k-anonymity as the privacy definition allowed suppression by default. This suppression approach does not apply to l-diversity since the purpose of anonymization is to diversify the sensitive value distribution. Therefore, suppression is disabled within this version of Incognito.

D Experiment Subject

1) Time Efficiency: Execution time is one of the aspects that must be considered while processing anonymization. To compare the execution time between algorithms, the software monitored time from the start till the very end of the anonymization process

2) Utility Metric: Due to the insufficiency of standardized metrics, it is challenging to measure data utility. Some metrics as discussed in [47], [48] are suitable for utility measure of numeric data but not provide any mechanism to measure utility for categorical data.For our evaluation criteria, the level of information remained in dataset after the completion of the anonymization process that is measure by using string metric proposed in [32]. In this analysis, 1.0 is considered as the best utility score, whereas zero measured as the worst score for maintaining utility. The Levenshtein metric measures the similarity between two words by calculating an edit distance. The distance is the number of deletions, insertions, or substitutions required to transform the anonymized data to original data. The procedure for calculating the Levenshtein distance between two strings X of length m and Y of length n is to calculate step by step in a matrix of order m x n edit distance between different sub-strings of X and Y. The corresponding values are stored in the matrix up to the box (m,n) which expresses the minimum distance between X and Y.

3) Privacy Metric: Attacker aims to re-identify anonymized data by linking with an individual's data record. The protection model applied in software used to measure identity disclosure risk by applying Shannon's entropy [34]. In this analysis, 1.0 is considered as best utility score, whereas zero is measured as the worst score for maintaining utility. Shannon's entropy metric was proposed in [35] as a measure of effective anonymity set. In order to measure identity disclosure risk of an anonymized dataset, Shannon entropy metric calculates the probability for the occurrence of values in each attribute. Usually, the more uncertain the probability, the less the disclosure risk. Shannon's entropy can be used to estimate this uncertainty, by applying:

$$H(X) = -\sum_{i=1}^{n} P(\, x\rrbracket_{i}\llbracket \square)log_{b}\llbracket P(x\rrbracket_{i}) \qquad (2)$$

where is a calculated probability of an attribute, compute it as the proportion of attribute in the dataset. The above equation quantifies the degree of utility contained in the dataset. The higher the entropy value, the less the disclosure risk.

4 Experiment's Results

For our experimental analysis we used following dataset:

- Employee's Salary dataset that consists of around 2000 records and 6 attributes (Name, Telephone, Age, Sex & Salary).
- Crime dataset that consists of around 1100 records and 7 attributes (Name, Block, Gender, Race, Age, Case Number and Cause of Incident).
- Marriage dataset that consists of around 800 records and 5 attributes (Country, Sex, Education, Marital Status & Age at Marriage).
- Disease dataset that consists of around 2000 records and 6 attributes (Name, Telephone, Age, Sex, Marital Status & Disease).
- Energy Consumption Dataset that consists of around 2000 records and 4 attributes (Residence Address, Zip code, Occupation & Energy Consumed).

In this section, we present the result generated from the datasets by presenting the comparison methodology.

A Employee Salary Dataset

Figure 4 shows that the algorithms perform efficiently with respect to the execution time. However, the execution time for l-Incognito is comparatively high.

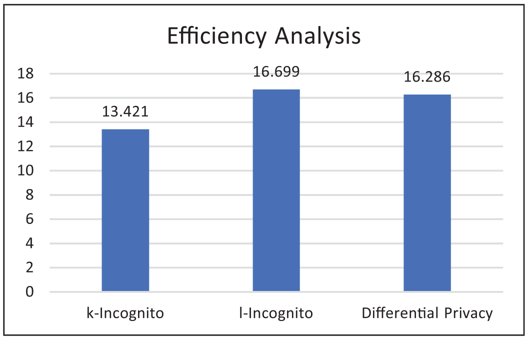


Figure 4: Efficiency Analysis of Employee Salary Dataset.

From Figure 5, it is obvious that K-Incognito performs better in maintaining utility. On the other hand, ε-differential privacy scored worst as it uses an additive noise mechanism for the anonymization process.

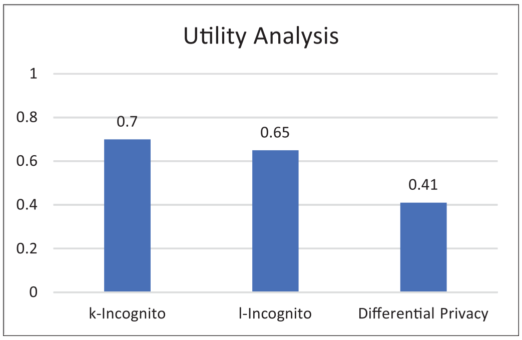


Figure 5: Utility Analysis of Employee Salary Dataset.

Figure 6 presents the result related to privacy analysis of the privacy models. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

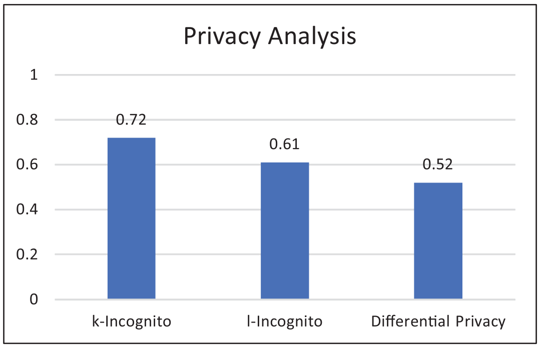


Figure 6: Privacy Analysis of Employee Salary Dataset.

B Patient Disease Dataset

Figure 7 shows that the execution time for l-Incognito scored the worst where as, k-Incognito complete the whole process is little less time as compared to the other algorithms.

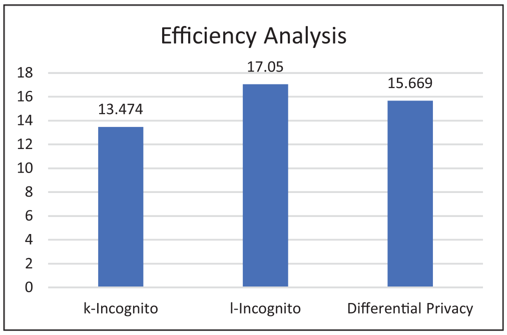


Figure 7: Efficiency Analysis of Patient Disease Dataset.

As in Figure 8, it is observed that again K-Incognito performed better in maintaining utility and differential privacy scored worst to gain utility of anonymized data.

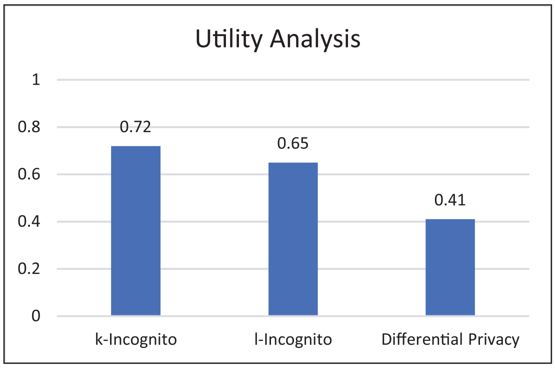


Figure 8: Utility Analysis of Patient Disease Dataset.

Figure 9 presents the privacy score of the anonymized dataset. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

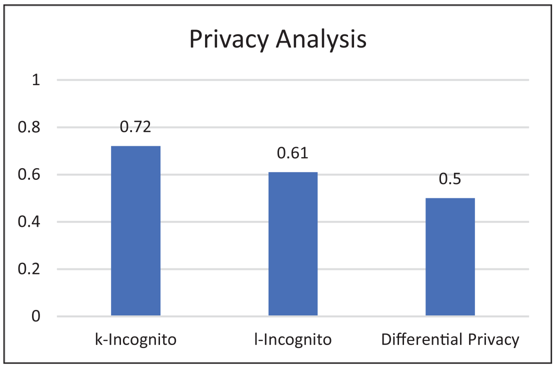


Figure 9: Privacy Analysis of Patient Disease Dataset.

C Crime Incident Dataset

Figure 10 shows that the algorithms perform the process in much less time as the number of quasi-attributes is less. The execution time for l-Incognito scores best while ε-differential privacy takes more time to complete anonymization.

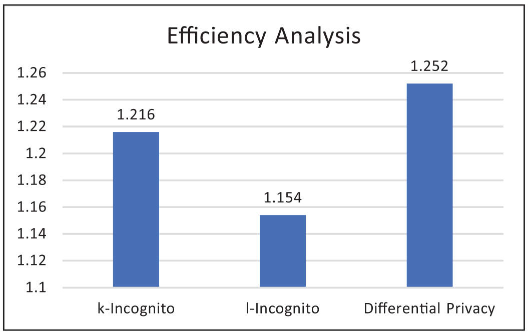


Figure 10: Efficiency Analysis of Crime Incident Dataset.

Figure 11, shows that K-Incognito scored best in maintaining utility. On the other hand, ε-differential privacy scored worst because of its additive noise mechanism for the anonymization.

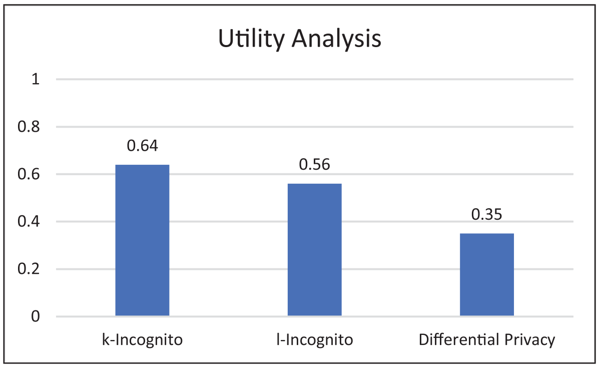


Figure 11: Utility Analysis of Crime Incident Dataset.

Figure 12 gives the result of privacy analysis of Crime Incident Dataset. K-Incognito reduces the risk disclosure and as the information is anonymized enough; it provides higher data privacy results.

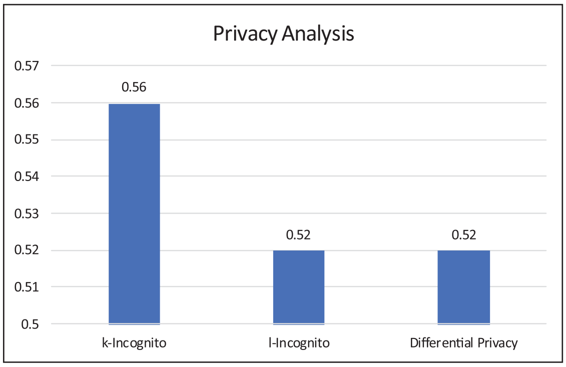


Figure 12: Privacy Analysis of Crime Incident Dataset.

D Marriage Dataset

Figure 13 shows that ε-differential privacy takes much more time while other algorithms perform the whole process in much less time. K-Incognito scored the best with respect to execution time.

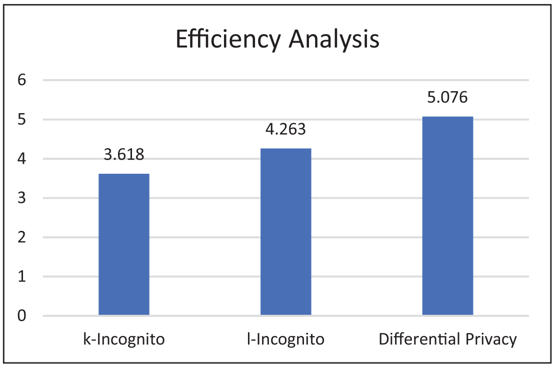


Figure 13: Efficiency Analysis of Marriage Dataset.

From Figure 14, can be found that K-Incognito performed most efficiently in maintaining the utility of dataset whereas, ε-differential privacy scored worst.

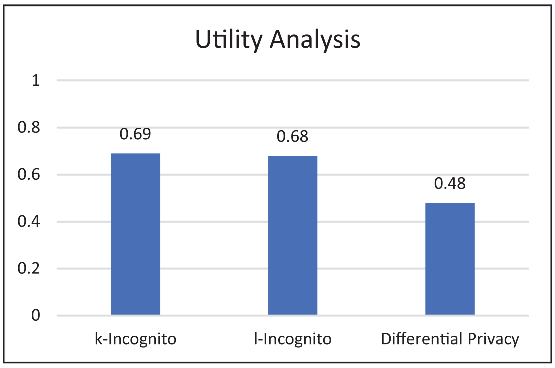


Figure 14: Utility Analysis of Marriage Dataset.

Figure 15 shows that the algorithms failed to provide the best privacy. l-Incognito provides better results related to privacy. However, k-Incognito is the worst performer when it comes to Marriage dataset.

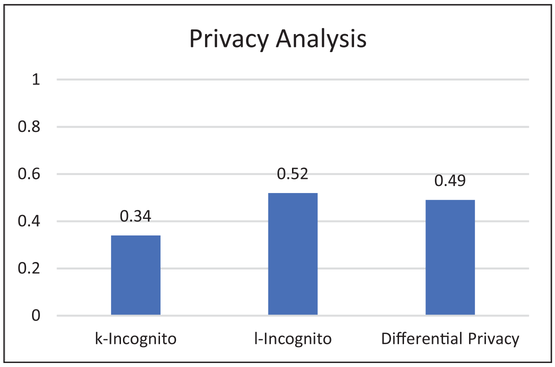


Figure 15: Privacy Analysis of Marriage Dataset.

E Energy Consumption Dataset

As this data consists of only 2 quasi identifiers, the algorithm takes a few seconds to anonymize the data. Figure 16 shows that the execution time for l-Incognito is comparatively less as compared to that of other algorithms.

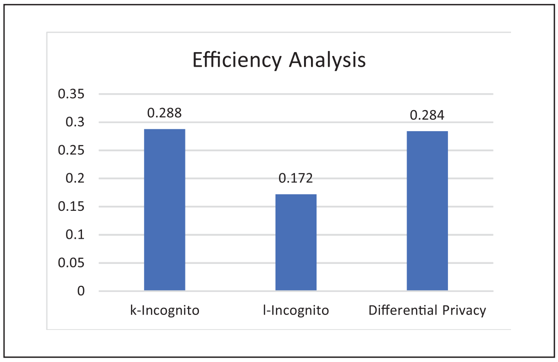


Figure 16: Efficiency Analysis of Energy Consumption Dataset.

From Figure 17, it is obvious that K-Incognito maintains better utility. Whereas, ε-differential privacy and l-Incognito provide less utility for anonymized data.

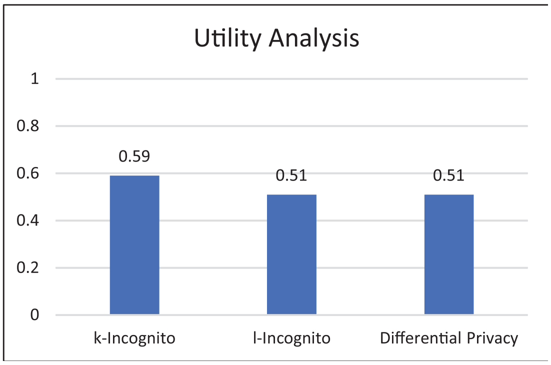


Figure 17: Utility Analysis of Energy Consumption Dataset.

Figure 18 presents the result related to privacy analysis of the privacy models. K-Incognito leads to minimize the risk disclosure and as the information is anonymized enough, it provides higher data privacy results.

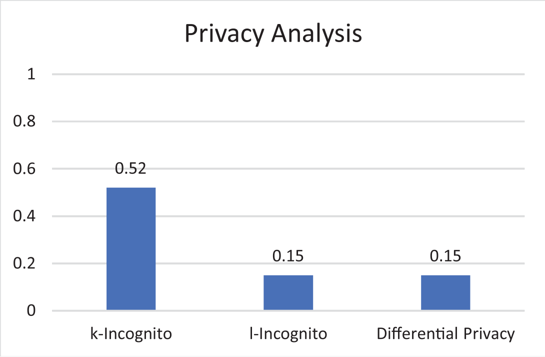


Figure 18: Privacy Analysis of Energy Consumption Dataset.

5 Conclusion

In this paper, we developed a data-publishing tool that preserves both the privacy and utility of data before sharing data to the external world. The software implemented different privacy model, including k-Incognito, l-Incognito and differential privacy. It evaluated the performance of algorithm using datasets related to different domains and compared the results in terms of execution time, data utility and data privacy. The software also features an automatic anonymization mode that provides results based on the algorithm that maintained utility-privacy of data at a higher rate. As per future enhancement, the tool can implement other algorithms. We can also introduce other supported file formats and work on the parameters i.e. de-identification risk of anonymized data.

Using k-Incognito, l-Incognito and differential privacy, we come to this conclusion that the execution time of algorithm depends upon the no. of quasi-identifiers contained by the dataset. The algorithm performs more efficiently when dataset consists of a smaller number of quasi-identifier. Moving forward, k-Incognito performs better in maintaining both utility and privacy in most of the cases. However, l-Incognito provides better privacy results when the dataset contains discrete values.

References

[1] U. Trivellato, “Microdata for Social Sciences and Policy Evaluation as a Public Good,” no. 11092, 2017.

[2] K. Wang, Privacy Preserving Data Publishing, vol. 42file:///, no. 4. 2010.

[3] L. Candela, D. Castelli, P. Manghi, and S. Callaghan, “On research data publishing,” Int. J. Digit. Libr., vol. 18, no. 2, pp. 73–75, 2017.

[4] A. Narayanan and V. Shmatikov, “How To Break Anonymity of the Netflix Prize Dataset,” 2006.

[5] O. Gkountouna, “A Survey on Privacy Preservation Methods,” pp. 1–30, 2011.

[6] G. Aggarwal et al., “Achieving anonymity via clustering,” ACM Trans. Algorithms, vol. 6, no. 3, pp. 1–19, 2010.

[7] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, “k-Anonymous data mining: a survey,” Adv. Database Syst., pp. 105–136, 2008.

[8] H. R. Asmaa and B. M. Y. Norizan, “Privacy preserving data publishing: Review,” Int. J. Phys. Sci., vol. 10, no. 7, pp. 239–247, 2015.

[9] L. Sweeney, “Simple demographics often identify people uniquely,” Carnegie Mellon Univ. Data Priv. Work. Pap. 3. Pittsburgh 2000, pp. 1–34, 2000.

[10] D. Dubli and D. K. Yadav, “Secure Techniques of Data Anonymization for Privacy Preservation,” vol. 8, no. 5, pp. 2015–2018, 2017.

[11] G. S and V. P, “A Survey on Privacy Preserving Data Publishing,” Int. J. Cybern. Informatics, vol. 3, no. 1, pp. 1–8, 2014.

[12] D. O. F. Informatics, “Survey of Privacy-Preserving Data Publishing Methods and Speedy : a multi-threaded algorithm preserving ? -anonymity Βιβλιογραφική επισκόπηση μεθόδων προστασίας της ιδιωτικότητας δεδομένων προς δημοσίευση και Speedy : ένας πολυνηματικός αλγόριθμος που δι,” no. October, 2015.

[13] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing,” ACM Comput. Surv., vol. 42, no. 4, pp. 1–53, 2010.

[14] Y. Xu, T. Ma, M. Tang, and W. Tian, “A survey of privacy preserving data publishing using generalization and suppression,” Appl. Math. Inf. Sci., vol. 8, no. 3, pp. 1103–1116, 2014.

[15] “GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES Published 25 January 2018,” no. January, 2018.

[16] A. Antoniades et al., “The effects of applying cell-suppression and perturbation to aggregated genetic data,” IEEE 12th Int. Conf. Bioinforma. Bioeng. BIBE 2012, no. November, pp. 644–649, 2012.

[17] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach to Privacy Preserving Data Publishing," IEEE Trans. Knowl. Data Eng., vol. PP, no. 99, p. 1, 2009.

[18] L. SWEENEY, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, pp. 557–570, 2002.

[19] L. Sweeney, "Datafly: a system for providing anonymity in medical data," pp. 356–381, 1998.

[20] Pierangela Samarati, "Protecting respondents' identities in micro- data release," IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010–1027, 2001.

[21] L.Sweeney,"ACHIEVINGfc-ANONYMITYPRIVACYPROTECTIONUSINGGENERALIZATION AND SUPPRESSION," Int. J. Uncertain., vol. 10, no. 5, pp. 571–588, 2002.

[22] P. Kulasinghe and A. El-Amawy, "On the Complexity of Optimal Bused Interconnections," IEEE Trans. Comput., vol. 44, no. 10, pp. 1248–1251, 1995.

[23] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowl. Data Eng., vol. 19, no. 5, pp. 1–14, 2007.

[24] G. Aggarwal et al., "Anonymizing tables," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3363 LNCS, pp. 246–258, 2005.

[25] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," Proc. - Int. Conf. Data Eng., vol. 2006, p. 25, 2006.

[26] K. Lefevre and D. J. Dewitt, "Workload-Aware Anonymization," pp. 277–286, 2006.

[27] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '06, p. 785, 2006.

[28] J. Gehrke, "ℓ -Diversity : Privacy Beyond k -Anonymity," vol. V, pp. 1–47, 2000.

[29] M. Ercan Nergiz, M. Atzori, and C. W. Clifton, Hiding the Presence of Individuals from Shared Databases. 2007.

[30] A. Paul Singh and D. Parihar Asst, "A Review of Privacy Preserving Data Publishing Technique," 2013.

[31] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and-Diversity."

[32] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness," IEEE Trans. Dependable Secur. Comput., pp. 1–1, 2017.

[33] J. Li, Y. Tao, and X. Xiao, Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. 2008.

[34] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439–450, Jun. 2000.

[35] X. Xiao, Y. Tao, M. Chen, and A. Kumar Maji, "Optimal Random Perturbation at Multiple Privacy Levels."

[36] R. Chaytor and K. Wang, “Small Domain Randomization: Same Privacy, More Utility,” 2150.

[37] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis.”

[38] V. S. Iyengar, “Transforming data to satisfy privacy constraints,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’02, 2002, p. 279.

[39] R. Dewri, I. Ray, I. Ray, and D. Whitley, “k-Anonymization in the Presence of Publisher Preferences,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 11, pp. 1678–1690, Nov. 2011.

[40] K. LeFevre, D. J. D. J. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain K-anonymity,” SIGMOD ’05 Proc. 2005 ACM SIGMOD Int. Conf. Manag. data, pp. 49–60, 2005.

[41] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “ℓ-Diversity: Privacy beyond k-anonymity,” Proc. - Int. Conf. Data Eng., vol. 2006, p. 24, 2006.

[42] C. Dwork, “Differential Privacy.”

[43] M. Tabata, Y. Hosokawa, O. Watanabe, and J. Sohma, “Direct Evidence for Main Chain Scissions of Polymers in Solution Caused by High Speed Stirring,” Polym J, vol. 18, no. 10, pp. 699–712, 1986.

[44] “Shannon entropy.”

[45] M. S. Simi, K. S. Nayaki, and M. S. Elayidom, “An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity,” IOP Conf. Ser. Mater. Sci. Eng., vol. 225, no. 1, 2017.

[46] C. Dwork, A. Roth, C. Dwork, and A. Roth, “The Algorithmic Foundations of Differential Privacy,” Found. Trends R □ Theor. Comput. Sci., vol. 9, pp. 211–407, 2014.

[47] D. Sinwar, R. Kaushik, and M. Tech Scholar, “Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering,” www.ijraset.com, vol. 2, 2014.

[48] F. Rahutomo, T. Kitasuka, and M. Aritsugi, “Semantic Cosine Similarity.”

Spreadsheet Based Software Engineering

Awais Azam[1] Khubaib Amjad Alam[2]

Abstract

Spreadsheets play a vital role in data processing and reporting procedures of any organization. That is why spreadsheet programming is the most successful end-user programming. Keeping in mind the similarity between spreadsheet programs and traditional programs the techniques that can be applied to traditional programs can also be applied in context of spreadsheets. The main objectives of this research are: (1) to classify spreadsheet research papers according to three criteria: techniques used, datasets used, publication channels and trends; and (2) to analyze these studies from four perspectives: study objectives, methods, method accuracy and limitations of the study. We perform a systematic mapping study on spreadsheet studies published in the period 2013-2018, collected from automated four electronic databases. We identified a total of 44 studies published between 2013 and 2018 and classified them on predefined classification criteria. Based on the findings of this research, it is concluded that Smell Detection is the technique that is applied in most of the cases on spreadsheets. The year 2016 receives the highest number of publications on spreadsheets. The most used dataset is from the corporation called EUSES.

Keyword: Software Engineering, Systematic Mapping, Spreadsheets, Spreadsheet Programming

1 Introduction

Research in software maintenance has shown that many programs contain a significant amount of duplicated (cloned) code. Such cloned code is considered harmful for two reasons: (1) multiple, possibly unnecessary, duplicates of code increase maintenance costs and, (2) inconsistent changes to cloned code can create faults, hence, lead to incorrect program behavior [1].

Spreadsheets have been widely used for various business tasks, including data management, decision support, financial reporting and so on. It is estimated that 90% of desktops have Excel installed [2] and there were over 55 million users in the United State working with spreadsheets in 2012 [3]. It is also believed that the number of spreadsheet programmers is bigger than that of software programmers [3]. In most of the cases, the people who are responsible for developing and maintaining spreadsheets are end users and they are not familiar with software development practices, so errors are easily induced into spreadsheets during maintenance and updates [4].

There is no mapping study applied on spreadsheets previously that demands from the research community to provide an overview of the trends and techniques that are being followed by researchers and are suitable to be applied on spreadsheets. Another problem in the previous studies is that each study focuses on some specific method or technique and does not cover the entire domain. The main goals and purpose of this research is to introduce a broader

[1] *National University of Computer & Emerging Sciences, Islamabad | awaisazam08@gmail.com*
[2] *National University of Computer & Emerging Sciences, Islamabad | khubaib.amjad@nu.edu.pk*

and precise overview of almost all the most commonly used latest techniques on spreadsheets in the form of a Systematic Mapping. We followed the guidelines of Petersen [4].

The organization of the paper is as follows: The detailed steps and the structured strategy of Systematic Mapping is described in section 2. Section 3 contains the presentation and discussion of results. Finally, the conclusion and future directions are given in section 4. Section 5 contains the references.

2 Systematic Mapping

A systematic mapping was conducted by following the guidelines of Petersen [4] and the collected data is analyzed in an unbiased and structured fashion. The first and the basic step to start the process of systematic mapping was the formulation of protocol that was designed and structured by Awais Azam and reviewed by Dr. Khubaib Amjad Alam. Now the steps performed in systematic mapping are described in the next sections.

A Research Questions

The research questions that were formed to escort mapping study are shown in Table 1. The table shows the research question and the motivation behind these research questions in order to clarify what was the main reason to include that research question in this research process.

Table 1: Research Question

RQ #	Research Question	Motivation
RQ 1	What Software Engineering methods have been applied on spreadsheets?	To identify the areas which are under consideration by the research community in case of spreadsheets.
RQ 2	What are the datasets and sources from where these datasets are collected?	The main focus of this question is to identify the datasets and their sources which will depict the overall trend in selecting the datasets for spreadsheets.
RQ 3	What is overall research productivity in the field of software engineering in the context of spreadsheets?	The purpose of this question is to give an idea of the overall research that is done and currently going on in this field.

B Search Strategy

An automated search was performed that consists of the following steps.

Data Sources: In order to answer the research questions, an automated search was performed on the previously constructed terms on the four electronic databases. Table 2 shows the database and the online link to that database.

Table 2: Electronic Databases

Database Name	Link
IEEE Xplore Digital Library	http://ieeexplore.ieee.org/
ACM Digital Library	http://dl.acm.org/
Science Direct	http://sciencedirect.com/
Springer	http://link.springer.com/

The studies that were the part of this research activity were from a time span of 2013 to 2018. The digital libraries that were considered are IEEE, ACM, Science Direct, Springer and Google Scholar databases based on title, abstract, and keywords.

Search Process: In order to make sure that we were not leaving any related study, a two-stage search process was adopted.

Initial search stage: Here, we used the proposed search terms to search for primary candidate studies in the four electronic databases. The retrieved papers were grouped together to form a set of candidate papers.

Secondary search stage: In this step, we reviewed all the studies retrieved after title-based search where we read the abstracts of the remaining studies and based on the abstract the studies which were not relevant were excluded and the studies that passed this search qualified for the full-text reading.

C Study Selection Procedure

This step was designed to get the most relevant studies which were retrieved from four electronic databases in order to answer the research questions. The selection procedure consists of the following basic steps:

- Initial records
- Title based records
- Abstract based records
- Full article-based records

As we progressed by following these steps the irrelevant studies kept on excluding in each step. Table 3 gives an overview of the number of studies that were selected at different stages from different databases.

Table 3: Study Selection Procedure

Database Name	Initial Records	Title Based	Abstract Based	Full Text
ACM	174	71	29	24
IEEE Explore	321	73	27	16
Science Direct	5	5	1	1
Springer	20	4	3	3

The study selection process can be visualized in detail and is represented in Table 3 that contains all the details of the number of studies at every stage and how much studies were excluded. The study selection process can be visualized in detail that is represented in Figure 1 that contains all the details of the number of studies at every stage and how much studies were excluded and the reason for exclusion and other related information can be found in Figure 1 that is constructed by following the PRISMA guidelines for systematic review.

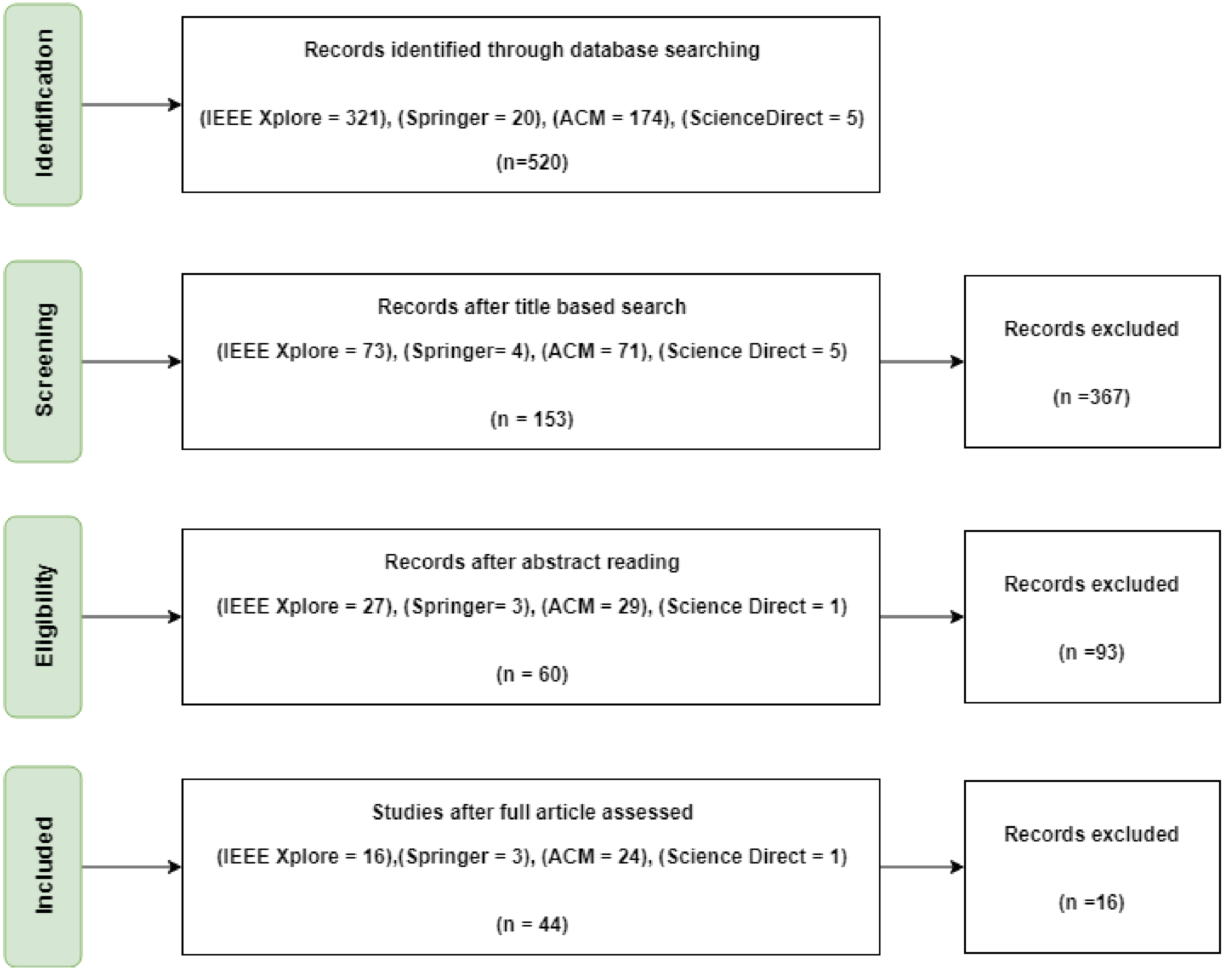
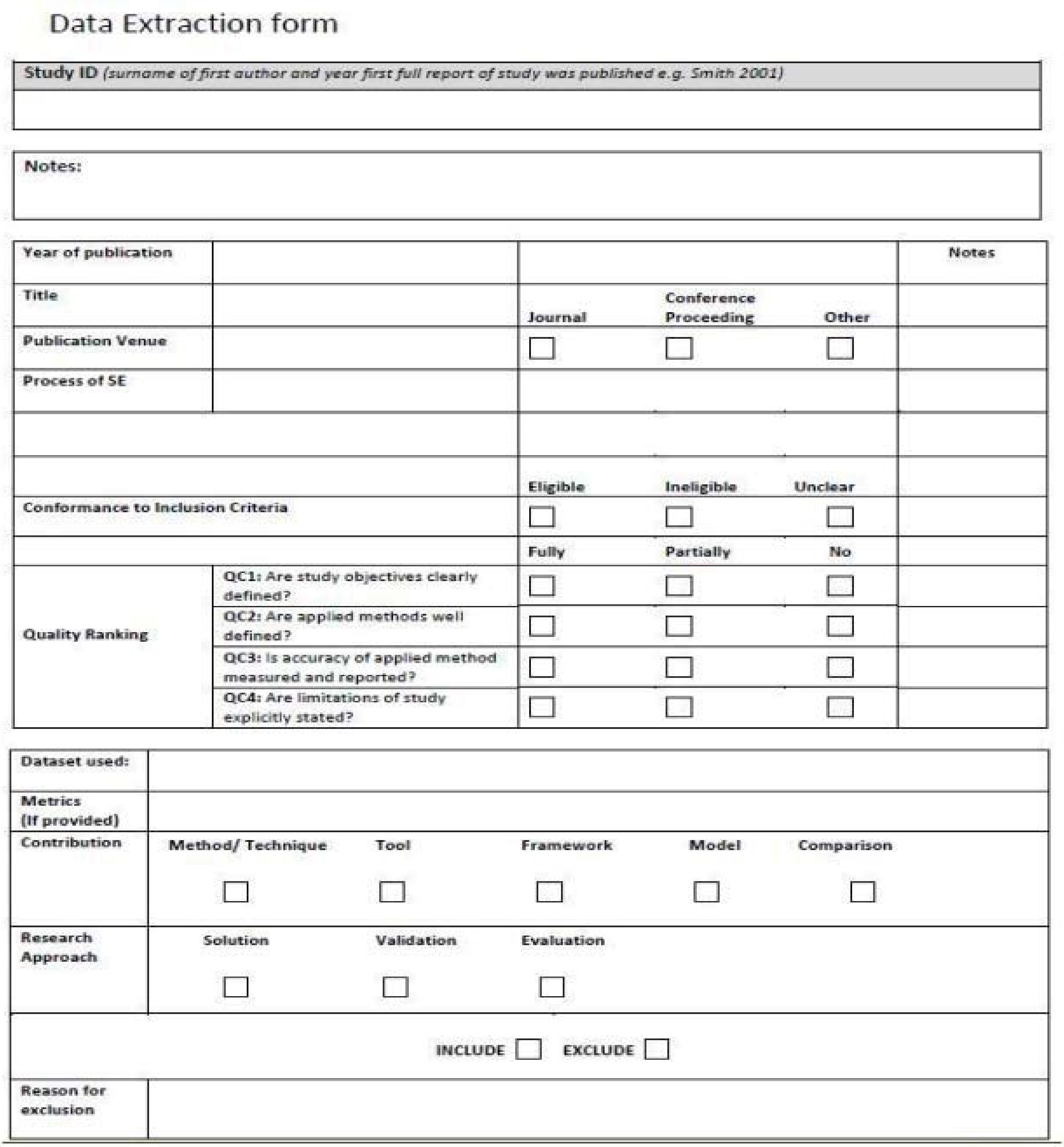

Identification
Records identified through database searching
(IEEE Xplore = 321), (Springer = 20), (ACM = 174), (ScienceDirect = 5)
(n=520)
Screening
Records after title based search
(IEEE Xplore = 73), (Springer= 4), (ACM = 71), (Science Direct = 5)
(n = 153)
Records excluded
(n =367)
Eligibility
Records after abstract reading
(IEEE Xplore = 27), (Springer= 3), (ACM = 29), (Science Direct = 1)
(n = 60)
Records excluded
(n =93)
Included
Studies after full article assessed
(IEEE Xplore = 16),(Springer = 3), (ACM = 24), (Science Direct = 1)
(n = 44)
Records excluded
(n =16)

Figure 1: PRISMA Flow Diagram

Data Extraction: The full text of all the qualifying studies was analyzed and the relevant information was extracted to already defined data extraction form shown in Figure 2.



Data Extraction form

Study ID *(surname of first author and year first full report of study was published e.g. Smith 2001)*

Notes:

Year of publication					**Notes**
Title		**Journal**	**Conference Proceeding**	**Other**	
Publication Venue		□	□	□	
Process of SE					
		Eligible	**Ineligible**	**Unclear**	
Conformance to Inclusion Criteria		□	□	□	
		Fully	**Partially**	**No**	
Quality Ranking	**QC1:** Are study objectives clearly defined?	□	□	□	
	QC2: Are applied methods well defined?	□	□	□	
	QC3: Is accuracy of applied method measured and reported?	□	□	□	
	QC4: Are limitations of study explicitly stated?	□	□	□	

Dataset used:					
Metrics (If provided)					
Contribution	**Method/ Technique** □	**Tool** □	**Framework** □	**Model** □	**Comparison** □
Research Approach	**Solution** □	**Validation** □	**Evaluation** □		
	INCLUDE □ **EXCLUDE** □				
Reason for exclusion					

Figure 2: Data Extraction Form

Inclusion and Exclusion Criteria: In this step, an inclusion and exclusion criteria were developed in order to further select the most related studies to carry out the research process. The inclusion and exclusion criteria are defined in Table 4.

Table 4: Inclusion and Exclusion Criteria

	Inclusion Criteria
IC1	Studies seeking convergence of software engineering in spreadsheets
IC2	Studies published in peer-reviewed conferences or journals
IC3	Studies published in or after 2013
IC4	Studies in English
	Exclusion Criteria
EC1	Studies with no validation of the proposed technique
EC2	Editorials, short papers, posters, technical reports, patents and reviews

Based on the criteria above, if the study meets the inclusion criteria and none of the exclusion criteria is met then such a study is further moved to the next stage that is quality assessment criteria.

Quality Assessment Criteria: This step was designed to ensure the quality of the studies that were finally going to be the part of the research process. There were total of four questions to estimate the quality of a study and each study is assessed against these four questions based on a 3-point scale. If the study answers the question it is indicated by Y (1 point) and if the study fails to answer the question it is indicated by N (0 points) and if it partially satisfies the answer then P (0.5 points) is given. The overall score required to include the study is 3 out of 4 to maintain high-quality standards. Table 5 contains the questions that were part of the quality assessment criteria.

Table 5: Quality Assessment Criteria

Question #	Criteria	Score
QC1	Are study objectives clearly defined?	Y|N|P
QC2	Are applied methods well defined?	Y|N|P
QC3	Is the accuracy of the applied method measured and reported?	Y|N|P
QC4	Are limitations of study explicitly stated?	Y|N|P

3 RESULTS

This section contains the results and discussion related to the research questions presented in Table. 1.

RQ1: Convergence of Software Engineering Methods with Spreadsheets

The graphical representation of the techniques and their usage in percentage is shown in Figure 3.

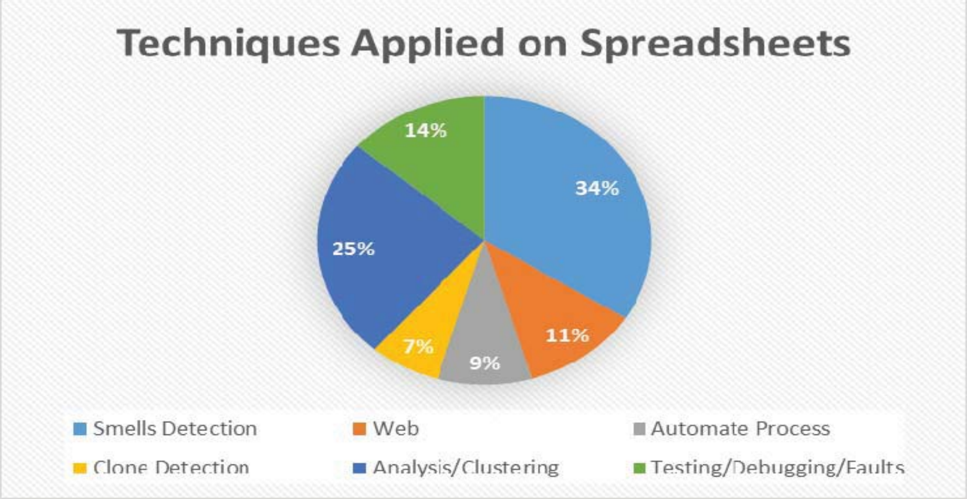


Figure 3: Graph of methods applied to Spreadsheets

It can be clearly seen in Figure 3 the highest percentage 34% is from smell detection which shows that the research community focused more on this area. The second place is acquired by Analysis/Clustering techniques which are 25%. It represents that this area is also the center of attention. Testing/Faults cover 14% of the overall studies. The area that is neglected by researchers is clone detection in spreadsheets that only occupies 7% of the overall research. Year wise distribution of studies with respect to areas on which they focused is shown in Figure 4. The studies are analyzed based on four areas. First is the Web, which means studies that are somehow related to the Web are covered in this area. Second is Automate Process, the studies that are involved in any kind of process that is automated using spreadsheets are included in this area. Third, is Testing/Debugging/Faults, the studies that are related to maintenance of spreadsheets are added in this area. The final and fourth area is Analysis/Comparison, in this area the studies in which any kind of analysis or comparison is done in the context of spreadsheets are included.

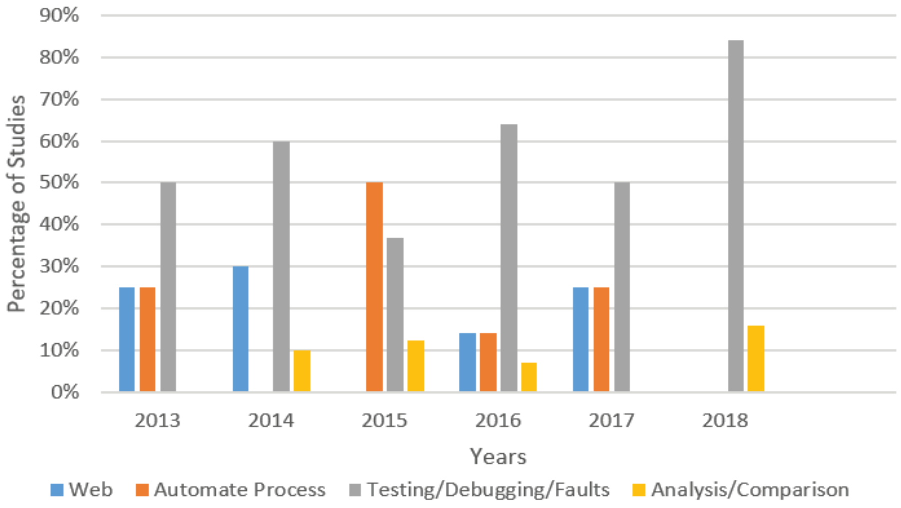


Figure 4: Year Wise Distribution of Studies

In case of year 2013, we can see that Web and Automate Process contains 25% of studies each and 50 % of studies in this year are focused on the research that is related to maintenance. In the year 2014, 30% of research is dedicated to the area of Web and a small portion of 10% is covered by Analysis/Comparison but the significant portion of this year is grabbed by the maintenance. The trend shows that Automate Process contains the highest number of studies in 2015 that is 50% and 37% of studies are included in the area of maintenance and around 12% of studies focused on Analysis/Comparison. It can be visualized from the chart that in the year 2016, all the four areas are covered by studies. Web and Automate Process covers 15% each. Maintenance got the highest percentage of 65% and Analysis/Comparison contributed by 7%. The year 2017 also represents the same trend that was observed in the year 2013. In 2018 85% of studies focused on maintenance and 15% of studies are for Analysis/Comparison. We can conclude from the above distribution of studies that the most focused area in the context of spreadsheets is maintenance from 2013 to 2018. However, it is also advocated that in every single year the highest inclination of research is towards the maintenance of spreadsheets.

RQ2: Data Sources and Datasets

The overall trend of the datasets that are currently preferred for spreadsheets can be visualized in Table 6.

Table 6: Data Sources

Data Source	No. of Studies
MS corpus	1
Enron	5
EUSES	11
Venron	1
DBpedia	1
Wiki Tables	1
ClueWeb09 Web crawl	1
F1F9	1
INFO1	1
Online Sources	29

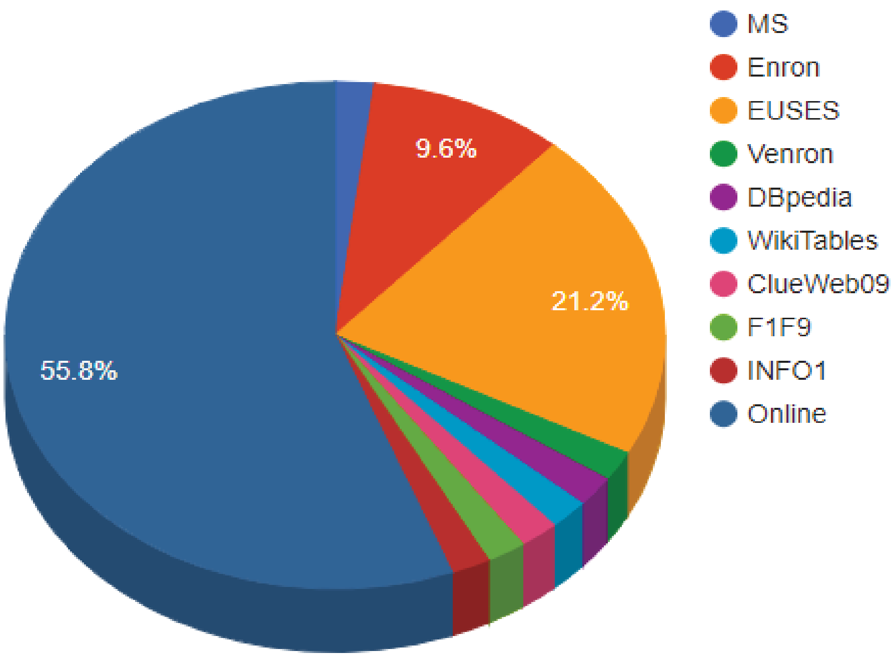


Figure 5: Data Sources

Figure 5 gives an overview in terms of the number of studies against the data sources. The highest number of studies suggests that most studies acquire data from online sources but it do not mean that such datasets do not belong to any of the data sources mentioned above, it means that the data is acquired from online and not from the official source. EUSES is also easily visible and is placed at the second position. Third position is for Enron corps. It can be concluded that the data sets that are mostly used to apply different methods on spreadsheets are from EUSES corps and Enron corps.

RQ3: Overall Research Productivity

Figure 6 shows the overall trend of the publications from the year 2013 to 2018. It can be observed that most studies are published in 2016. From the year 2013 to 2016, there is an increase in the number of studies and in 2017 and 2018, shows identical results.

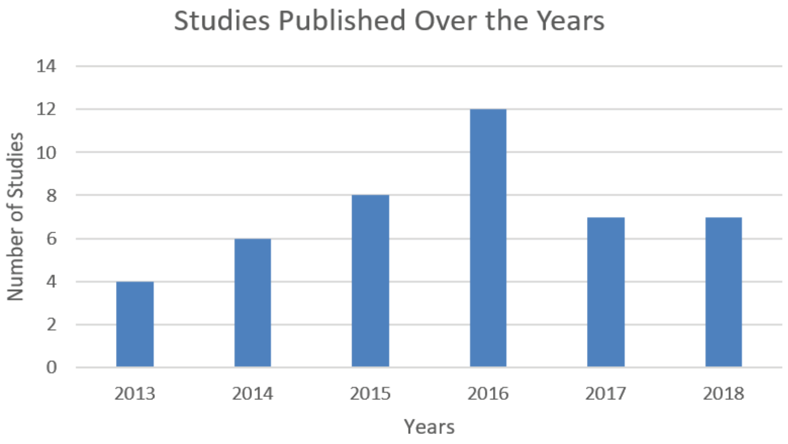


Figure 6: Studies Published over the Years

Table 7: Quality Levels of Selected Studies

Quality Level	Number of Studies	Percentage (%)
Very High (score = 4)	11	25 %
High (score = 3.5)	13	29.54 %
Medium (score = 3)	20	45.45 %
Total	44	100.00 %

Table 7 reveals that 25% of the studies lies in the highest quality span. While more than 29% of studies managed to qualify for the high score and the most number of studies are from medium quality which is around 45% according to our predefined quality assessment criteria. The research approaches that are considered when it comes to spreadsheets are of three types, solution, validation, evaluation.

Table 8: Research Approaches

Research Approach	Number of Studies	Percentage (%)
Solution	34	77.27 %
Validation	6	13.63 %
Evaluation	4	9.09 %
Total	44	100.00 %

Table 8 presents the trend of the approaches followed by studies to perform different methods on spreadsheets. Solution approach is most adopted by the researchers and covers about 77% of the overall studies.

Table 9 gives a detailed overview of the venues in which the research papers about spreadsheets are published. International Conference on Software Engineering is the most visible venue as most number of studies are published in it.

Table 9: Publication Venues

Venue	Number of Studies
International Conference on Software Engineering	6
Symposium on User interface software and technology	2
International Conference on Knowledge Discovery and Data Mining	2
International Conference on Software Analysis, Evolution and Re-engineering	2
Symposium on Visual Languages and Human-Centric Computing	2
International Conference on Software Maintenance and Evolution	2
IEEE TRANSACTIONS ON SOFTWARE ENGINEERING	3
International Conference on Inventive Computation Technologies	2

Automated Software Engineering	2
International Conference on Knowledge Capture	1
Software Engineering Conference and Symposium on the Foundations of Software Engineering	1
Proceedings of the ACM on Programming Languages	1
SIGPLAN Notices	1
International Symposium on Software Testing and Analysis	1
Brazilian Symposium on Systematic and Automated Software Testing	1
International Workshop on Semantic Search over the Web	1
Conference on Information and Knowledge Management	1
International Database Engineering & Applications Symposium	1
International Symposium on Foundations of Software Engineering	1
International Conference on Software Engineering Companion	1
Conference on Human Factors in Computing Systems	1
International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software	1
Conference of the Center for Advanced Studies on Collaborative Research	1
Conference on Programming Language Design and Implementation	1
International Conference on Mining Software Repositories	1
International Symposium on Software Reliability Engineering Workshops	1
IEEE Transactions on Reliability	1
International Workshop on Document Analysis Systems	1
Science of Computer Programming	1
Empirical Software Engineering	1

The contribution of studies that is published by the researchers depends on the nature of the research and the intention of the publisher. There are several types of contribution that is made to the research community. Some researchers do a comparison among different studies, some prefer to propose a new model, framework, tool or some kind of method/technique. In case of spreadsheets the detailed overview of the contribution can be found in Figure 7.

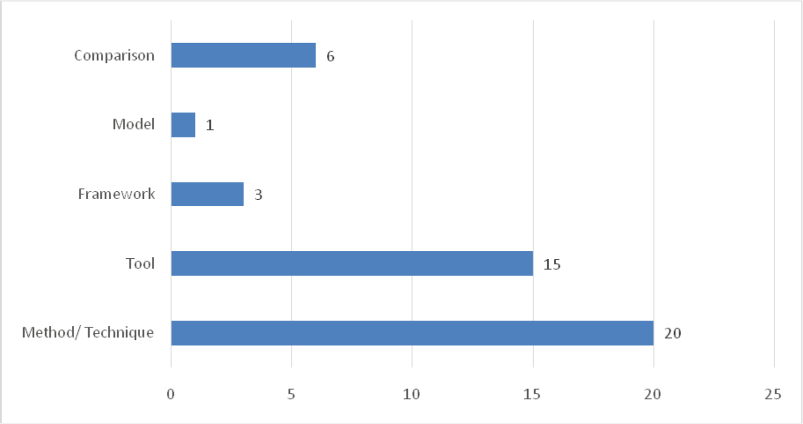


Figure 7: Contribution Facets

As the Figure 7 depicts the trend that 20 of the total studies published focuses on the presentation of some new technique/method. After that, 15 studies proposed a tool. Then 6 studies give a comparison of already existing techniques or tools. New framework was given in 3 publications. And only one study proposed a model.

Another interesting finding is that which countries are actively participating in the spreadsheet research and this can be found out by having a look at the Figure 8, which briefly shows the country name and the number of studies published in that country.

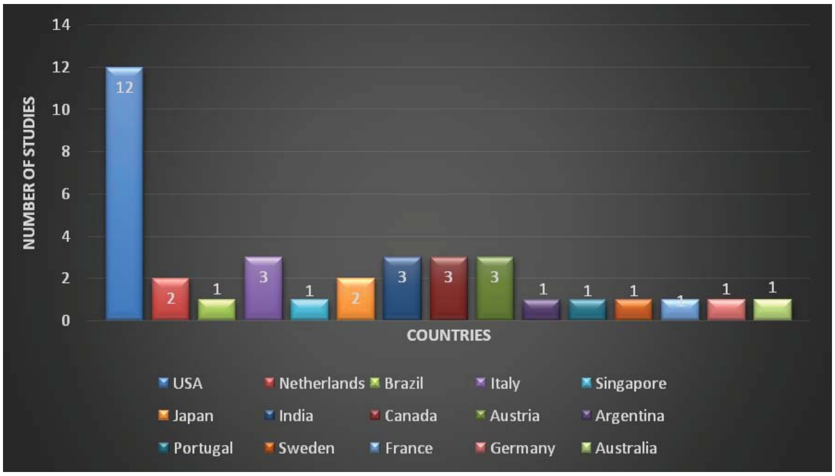


Figure 8: Contribution Facets

There are 12 studies that were published in USA, so USA has the biggest contribution towards spreadsheet research. After that second place is claimed by India, Canada and Austria containing 3 studies each. At third place there are two countries Netherlands and Japan with 2 studies published in each country.

4 CONCLUSIONS

This systematic mapping study summarizes the existing studies with their focus on the spreadsheets and the methods or processes that are applied on spreadsheets. The paper presents a range of papers on spreadsheets and classifies them according to different criteria like techniques, approaches, datasets and quality. The primary search fetched 520 studies. Then title-based search was performed to get the most relevant papers and 153 papers were extracted. After that, abstract reading was performed to further extend the filtration process and 60 papers managed that filter. Then these papers were passed through the predefined research questions and quality assessment criteria test and finally 44 studies were selected to carry out this systematic mapping study. The main findings of this research are as follows:

- Clone detection is the most neglected area in the context of spreadsheets.
- In most of the studies, researchers gave new techniques or they were able to improve already existing techniques which show solution research is the approach that gained more importance.
- The most used dataset for spreadsheets is from EUSES and Enron.

Software engineering methods application on spreadsheets is gaining importance with time because it is becoming more and more important for the research community because there is a high potential in this area for researchers. The main objective of research in spreadsheets is to get the most out of it because almost every institution, company or business is utilizing spreadsheets in one way or the other.

Acknowledgments

The authors would like to thank Usman Ali (PhD Scholar at Institute of Business Administration) for his valuable feedback on this paper.

References

[1] Juergens, E., Deissenboeck, F., Hummel, B., & Wagner, S. (2009, May). Do code clones matter?. In Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on (pp. 485-495). IEEE.

[2] Bradley, L., & McDaid, K. (2009, May). Using Bayesian statistical methods to determine the level of error in large spreadsheets. In Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on (pp. 351-354). IEEE.

[3] Scaffidi, C., Shaw, M., & Myers, B. (2005, September). Estimating the numbers of end users and end user programmers. In Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on (pp. 207-214). IEEE.

[4] Petersen, K., Vakkalanka, S., &Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1-18.

[5] de Vos, M., Wielemaker, J., Schreiber, G., Wielinga, B., & Top, J. (2015, October). A methodology for constructing the calculation model of scientific spreadsheets. In Proceedings of the 8th International Conference on Knowledge Capture (p. 2). ACM.

[6] Zhang, J., Han, S., Hao, D., Zhang, L., & Zhang, D. (2018, October). Automated refactoring of nested-if formulae in spreadsheets. In Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 833-838). ACM.

[7] Inala, J. P., & Singh, R. (2017). WebRelate: integrating web data with spreadsheets using examples. Proceedings of the ACM on Programming Languages, 2(POPL), 2.

[8] Almasi, M. M., Hemmati, H., Fraser, G., McMinn, P., &Benefelds, J. (2018, July). Search-based detection of deviation failures in the migration of legacy spreadsheet applications. In Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis(pp. 266-275). ACM.

[9] Almeida, L., Cirilo, E., & Barbosa, E. A. (2016, September). SS-BDD: Automated Acceptance Testing for Spreadsheets. In Proceedings of the 1st Brazilian Symposium on Systematic and Automated Software Testing (p. 5). ACM.

[10] Benson, E., Zhang, A. X., & Karger, D. R. (2014, October). Spreadsheet driven web applications. In Proceedings of the 27th annual ACM symposium on User interface software and technology (pp. 97-106). ACM.

[11] Chang, K. S. P., & Myers, B. A. (2014, October). Creating interactive web data applications with spreadsheets. In Proceedings of the 27th annual ACM symposium on User interface software and technology (pp. 87-96). ACM.

[12] Barowy, D. W., Gochev, D., & Berger, E. D. (2014, October). CheckCell: data debugging for spreadsheets. In ACM SIGPLAN Notices (Vol. 49, No. 10, pp. 507-523). ACM.

[13] Chen, Z., &Cafarella, M. (2013, August). Automatic web spreadsheet data extraction. In Proceedings of the 3rd International Workshop on Semantic Search over the Web (p. 1). ACM.

[14] Chen, Z., Dadiomov, S., Wesley, R., Xiao, G., Cory, D., Cafarella, M., & Mackinlay, J. (2017, November). Spreadsheet property detection with rule-assisted active learning. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 999-1008). ACM.

[15] Cheung, S. C., Chen, W., Liu, Y., & Xu, C. (2016, May). CUSTODES: automatic spreadsheet cell clustering and smell detection using strong and weak features. In Proceedings of the 38th International Conference on Software Engineering(pp. 464-475). ACM.

[16] Compton, M., Ratcliffe, D., Shu, Y., & Squire, G. (2015, July). Declarative Data Exchange: Spreadsheets to RDF/OWL. In Proceedings of the 19th International Database Engineering & Applications Symposium (pp. 96-105). ACM.

[17] Dou, W., Cheung, S. C., & Wei, J. (2014, May). Is spreadsheet ambiguity harmful? detecting and repairing spreadsheet smells due to ambiguous computation. In Proceedings of the 36th International Conference on Software Engineering (pp. 848-858). ACM.

[18] Dou, W., Cheung, S. C., Gao, C., Xu, C., Xu, L., & Wei, J. (2016, November). Detecting table clones and smells in spreadsheets. In Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (pp. 787-798). ACM.

[19] Dou, W., Xu, L., Cheung, S. C., Gao, C., Wei, J., & Huang, T. (2016, May). VEnron: a versioned spreadsheet corpus and related evolution analysis. In Software Engineering Companion (ICSE-C), IEEE/ACM International Conference on(pp. 162- 171). IEEE.

[20] Chang, K. S. P., & Myers, B. A. (2016, May). Using and exploring hierarchical data in spreadsheets. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2497-2507). ACM.

[21] Hermans, F., Sedee, B., Pinzger, M., & van Deursen, A. (2013, May). Data clone detection and visualization in spreadsheets. In Software Engineering (ICSE), 2013 35th International Conference on (pp. 292-301). IEEE.

[22] Hermans, F., & Murphy-Hill, E. (2015, May). Enron's spreadsheets and related emails: A dataset and analysis. In Proceedings of the 37th International Conference on Software Engineering-Volume 2 (pp. 7-16). IEEE Press.

[23] Li, K., He, Y., &Ganjam, K. (2017, August). Discovering enterprise concepts using spreadsheet tables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1873-1882). ACM.

[24] Koch, P., Schekotihin, K., Jannach, D., Hofer, B., Wotawa, F., & Schmitz, T. (2018, May). Combining spreadsheet smells for improved fault prediction. In Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results (pp. 25-28). ACM.

[25] Kovalenko, O., Serral, E., &Biffl, S. (2013, September). Towards evaluation and comparison of tools for ontology population from spreadsheet data. In Proceedings of the 9th International Conference on Semantic Systems (pp. 57-64). ACM.

[26] McCutchen, M., Itzhaky, S., & Jackson, D. (2016, October). Object spreadsheets: a new computational model for end-user development of data-centric web applications. In Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (pp. 112-127). ACM.

[27] Hermans, F. (2013, November). Improving spreadsheet test practices. In Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (pp. 56-69). IBM Corp..

[28] Barowy, D. W., Gulwani, S., Hart, T., & Zorn, B. (2015, June). FlashRelate: extracting relational data from semi-structured spreadsheets using examples. In ACM SIGPLAN Notices (Vol. 50, No. 6, pp. 218-228). ACM.

[29] Xu, L., Dou, W., Gao, C., Wang, J., Wei, J., Zhong, H., & Huang, T. (2017, May). SpreadCluster: recovering versioned spreadsheets through similarity-based clustering. In Proceedings of the 14th International Conference on Mining Software Repositories (pp. 158-169). IEEE Press.

[30] Ayalew, Y., Clermont, M., &Mittermeir, R. T. (2008). Detecting errors in spreadsheets. arXiv preprint arXiv:0805.1740.

[31] Koch, P., &Schekotihin, K. (2018, October). Fritz: A Tool for Spreadsheet Quality Assurance. In 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)(pp. 285-286). IEEE.

[32] Pandita, R., Parnin, C., Hermans, F., & Murphy-Hill, E. (2018, October). No half-measures: A study of manual and tool-assisted end-user programming tasks in Excel. In 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC) (pp. 95-103). IEEE.

[33] Koch, P., Schekotihin, K., Jannach, D., Hofer, B., Wotawa, F., & Schmitz, T. (2018, May). Combining spreadsheet smells for improved fault prediction. In Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results (pp. 25-28). ACM.

[34] Abreu, R., Cunha, J., Fernandes, J. P., Martins, P., Perez, A., & Saraiva, J. (2014, September). Faultysheet detective: When smells meet fault localization. In Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on(pp. 625-628). IEEE.

[35] Abreu, R., Cunha, J., Fernandes, J. P., Martins, P., Perez, A., & Saraiva, J. (2014, September). Smelling faults in spreadsheets. In Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on (pp. 111-120). IEEE.

[36] Cunha, J., Fernandes, J. P., Mendes, J., & Saraiva, J. (2015). Embedding, evolution, and validation of model-driven spreadsheets. IEEE Transactions on Software Engineering, 41(3), 241-263.

[37] Dou, W., Xu, C., Cheung, S. C., & Wei, J. (2017). CACheck: Detecting and Repairing Cell Arrays in Spreadsheets. IEEE Trans. Software Eng., 43(3), 226-251.

[38] Hermans, F., & Murphy-Hill, E. (2015, May). Enron's spreadsheets and related emails: A dataset and analysis. In Proceedings of the 37th International Conference on Software Engineering-Volume 2 (pp. 7-16). IEEE Press.

[39] Hermans, F., Jansen, B., Roy, S., Aivaloglou, E., Swidan, A., &Hoepelman, D. (2016, March). Spreadsheets are code: An overview of software engineering approaches applied to spreadsheets. In Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on (Vol. 5, pp. 56-65). IEEE.

[40] Hofer, B. (2017, October). Removing Coincidental Correctness in Spectrum-Based Fault Localization for Circuit and Spreadsheet Debugging. In Software Reliability Engineering Workshops (ISSREW), 2017 IEEE International Symposium on (pp. 199-206). IEEE.

[41] Hofer, B., Höfler, A., &Wotawa, F. (2017). Combining models for improved fault localization in spreadsheets. IEEE Transactions on Reliability, 66(1), 38-53.

[42] Jansen, B., &Hermans, F. (2015, September). Code smells in spreadsheet formulas revisited on an industrial dataset. In Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on (pp. 372-380). IEEE.

[43] Koci, E., Thiele, M., Lehner, W., & Romero, O. (2018, April). Table Recognition in Spreadsheets via a Graph Representation. In 2018 13th IAPR International Workshop on Document Analysis Systems (DAS) (pp. 139-144). IEEE.

[44] Macedo, N., Pacheco, H., Sousa, N. R., & Cunha, A. (2014, July). Bidirectional spreadsheet formulas. In Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on (pp. 161-168). IEEE.

[45] Rajdev, U., & Kaur, A. (2016, August). Automatic detection of bad smells from excel sheets and refactor for performance improvement. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 2, pp. 1-8). IEEE.

[46] Rajdev, U., & Kaur, A. (2016, August). Bad smell refactor removable detection of audit spreadsheet. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 1, pp. 1-7). IEEE.

[47] Cunha, J., Mendes, J., Saraiva, J., & Visser, J. (2014). Model-based programming environments for spreadsheets. Science of Computer Programming, 96, 254-275.

[48] Hermans, F., Pinzger, M., & van Deursen, A. (2015). Detecting and refactoring code smells in spreadsheet formulas. Empirical Software Engineering, 20(2), 549-575.

[49] Jannach, D., & Schmitz, T. (2016). Model-based diagnosis of spreadsheet programs: a constraint-based debugging approach. Automated Software Engineering, 23(1), 105-144.

[50] Schmitz, T., Hofer, B., Jannach, D., &Wotawa, F. (2016, July). Fragment-based diagnosis of spreadsheets. In Federation of International Conferences on Software.

Robust Feature Extraction Techniques in Speech Recognition: A Comparative Analysis

Isra Khan[1] Ashhad Ullah[2] Rafi Ullah[3] Shah Muhammad Emad[4]

Abstract

As the world is moving towards new era known as the era of 'Artificial intelligence' where many of things will be controlled automatically through many sources such as face and thumb lock like this we can control things through sound as people are trying to do so and this thing getting hot day by day but it is not explored that much, in this paper we are exploring sound and its feature extraction techniques through which we can extract features from various types of sound and can make them applicable as this paper presents a survey on feature extraction to comparative analysis with respect to properties such as noisy data, complexity, accuracy, extraction method it will be helpful to use which data set with which type of sound. Feature extractions process has a direct relation with any of the machine learning algorithm. If feature extracted is robust, the use of underlining machine learning algorithm will increase accuracy. This paper targeted only the comparative analysis of features used in literature for sound. In future, two or more features will be combined to enhance the impact of sound recognition systems.

Keywords: Sound Recognition, Feature Extraction in Sound Recognition, Sound Detection, Robust Feature in Sound Recognition and Detection, Robust Features in speech recognition.

1 Introduction

Sound is the vibration that travels through air or any medium and these vibrations are audible when they reach an individual's ear and sound is formed by the unbroken and consistent vibrations. The first ever sound that was noted by invented by Édouard-Léon Scott de Martinville, was assembled by a gadget called a phonautograph in 1957. Phonautograph write out sound waves into a line that is drawn on paper but with these waves there are some features through which sound can be categorized in many classes or categories were extracted. Let's take an example when we hear any kind of sound our brain start processing on it and categorize that sound like we can predict that this is the voice of a female without seeing that female because we know which value range belongs to which category, but the major challenge is to extract the features and their different ways of doing it such as MFCC, RASTA, LPCC, Cepstral Analysis, LPC and many others [1]. The majority of these proposed frameworks consolidate two handling stages. The first stage studies the received sound wave and extracts parameters (features) from it. The feature extraction the extracted features and both of these stages are defined briefly below. Many set of feature extraction are proposed earlier for audio classification [2,3,4]. Largest portion has been covered by low-level signal features and then second important feature set consist of Mel-frequency cesptral coefficient (MFCC) [5] and then those remaining features come.

[1] *Karachi Institute of Economics & Technology, Karachi | khanisra@gmail.com*
[2] *Karachi Institute of Economics & Technology, Karachi | ashhadullah19@gmail.com*
[3] *Karachi Institute of Economics & Technology, Karachi | rafiafridi783@gmail.com*
[4] *Karachi Institute of Economics & Technology, Karachi | shahmuhammademad@gmail.com*

All of the features are used audio classification and are very powerful in classifying the audio class but it gradually decrease when amount of classes increase. Therefore, using which feature set with which amount of classes is an issue which can create further more issues if we select wrong feature set with respect to the problem description, result will help you with comparison done which will guide you when to use which set [6, 7].

Speech is that the most typical manner of communication between humans. Speech also carries the information related to the speaker. To recognize the speaker there are features exists in the speech signal. These extracted features will be useful in training of the model for speech recognition.

In audio processing, feature extraction is the backbone. The importance of feature extraction technique can never be ignored in speech recognition and processing systems [8]. But these features that are extracted must fulfill these criteria while doing speech recognition. These standards are [9]:

- Easy to measure extracted speech features
- Not be susceptible to mimicry
- Perfect in showing environment variation
- Stability over time

For feature extraction audio samples are collected and then converted to digital signals at a regular interval. At these voice samples noise reduction is performed so that the original audio sample can be find to perform feature extraction on it. For the speech recognition we extract the features from the digital signals which provide the acoustic properties of that specific digital dataset that is really useful for representing the speech signal.

These speech signals are slowly timed varying signals (quasi-stationary). When analyzed for a short time interval for example examined for example 5ms-100ms, the attributes seems to be relatively stationary. However if sound/vocal features are modified for a specified time interval, it reflects the different values of spoken audio features. The information of audio signal can be categorized by using short term amplitude spectrum of the audio wave form. These techniques are known as phonemes helps in the extraction of sound features of short term amplitude spectrum from audio signals called phonemes [10].

Rest of the paper is divided as follow; Section I is about the literature review or Related Work, Section II is the detail explanation of different features that can be extracted from sound, Section IV is Result section that is detail comparative analysis of different features extraction technique, Section V is the concluding the topic and Section VI is the future potential area.

2 Related Work

Authors of [11] focused on the comparative analysis of widely used feature extraction techniques related to speech recognition and in the end of the research has conclude that the PLP is extracted on the conception of logarithmically spaced filter bank, combined with the conception of human hearing system and has improved results than LPC.

According to paper [29], author has extracted MFCC feature and de-noise the audio sample and also enhanced the MFCC feature by calculating the delta energy for the coefficient.

Authors has extracted MFCC feature for the speech emotion detection discussed in detail in [30]. MFCC feature is extracted and worked very efficiently and train the model for the detecting of speech detection emotion.

Isolated speech recognition by using the MFCC and Dynamic Time Wrapping (DTW) was focused by the authors of [31]. In this research features for the isolated speech recognition were extracted by using the MFCC.

In [14], authors has identified and focused on the problem of optimizing the acoustic features set by Ant Colony Optimization for the Automatic speech recognition. Speech signal is considered as input in this research and feature extraction is performed over this signal using MFCC extraction method, total 39 coefficients are extracted in this research by using MFCC.

Comparative analysis of speech recognition has done in paper [33]. These analysis was performed on noisy conditions on the widely used feature extraction techniques named MFCC,LPCC,PLP, RASTA-PLP and HMM and has analyzed that PLP distinctly gave the maximum percentage of recognition and the grouping of LPCC, PLP and RASTA provided the output as third maximum recognition percentage.

In [34], Authors have worked on the change in detection in multi-dimensional unlabeled data in which features were extracted by using the PCA feature extraction technique.

According to the authors of [35], they focused on the PCA drawbacks which are high computational cost, extensive memory utilization and low adequacy in handling expansive dimensional datasets, so author has proposed a new technique Folded-PCA. By using this new proposed technique these drawbacks can be resolve.

Drawbacks of PCA was discussed in paper [36]. These drawbacks are: computational cost, extensive memory utilization and low adequacy in handling expansive dimensional datasets, so they analyzed two variation of the PCA technique SPCA and Seg-PCA. These variations can be helpful to reduce the drawbacks of PCA.

Authors in [20] have done the survey over the feature extraction technique and conclude that the LPC is vector dimension and has high computational cost and also reduce accuracy and their window size which is not good for non-stationary speech signals such as speech signal.

In [38], Authors has proposed the new technique for the noisy speech recognition based on auditory filter modeling-based feature extraction and gives the result that LPC is less efficient in this manner in comparison with PLPaGc.

Comparative analysis for the speech recognition specific for Hindi language words, and has analyzed that LPCC gives less recognition rate for isolate, paired and hybrid words as compared to MFCC has performed in [39].

A new recognition system was proposed in [40]. This system uses the acoustic waves generated by the construction equipment, this will be very helpful to avoid external damages. Feature extraction for the recognition system was done by combining LPCC and SVM.

The RASTA feature extraction technique in combination with TANDEM was used by the authors of [41]. The authors stated that this technique is an efficient way to represent the message-information in the speech signal.

3 Feature Extraction Techniques

Various Features Extraction techniques has been observed in the literature, that is used for sound recognition and sound detection. Each one has its own advantages and disadvantages depending upon the environment I-e nature of problem. For example features extraction used in sounds related to school cafe will be having different impact on sound of vehicles. Some of the features extraction techniques that are observed during our research are:

- Mel-frequency Cepstral Coefficients (MFCC)
- Perceptual Linear Predictive (PLP)
- Relative Spectral Processing (RASTA)
- Linear Prediction Cepstral Coefficient (LPCC)
- Principle Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Wavelet
- Dynamic Time Warping (DTW)
- Combined LPC and MFCC
- Kernel based feature extraction
- Independent Component Analysis (ICA)
- Integrated phoneme subspace method
- Probabilistic Linear Discriminant Analysis (PLDA)
- Linear Prediction Coefficient (LPC)
- Discrete Wavelet Transformation (DWT)
- Wavelet Packet Decomposition (WPD)
- Gammatone Frequency Cepstral Coefficient (GFCC)
- Gaussian Mixture Model (GMM)

This paper targeted only six features extraction technique (MFCC, PLP, PCA, LPC, LPCC and RASTA) to compare on the basis of several parameters such as Impact in presence of noise I-e noisy data, Complexity (in case of features extraction and computation), Accuracy and Feature Extraction Method.

A Mel-Frequency Cepstral Cofficients

MFCC is one of the most important techniques used to extract the feature from speech signal [11] that is actually based over the human's ear scale bandwidth. It uses the low and the high frequencies, measured in Hertz (Hz) to get the speech signal. MFCCs considered as frequency domain features that are more accurate in comparison with time domain features [11]. These signals are then divided into the audio frames to calculate the MFCC. Let each frame of audio signal contains the N samples and considers the next and the previous frames of the audio signal is separated by M samples where M<N. All audio frames are multiplied by a Hamming window. The hamming window [16] value can be calculated using this equation 1.

$$W(n) = 0.54 - 0.46\, cos(2^{\pi n}/_{N} - 1) \qquad (1)$$

Then speech signal is transformed to frequency domain from time domain by utilizing its Discrete Fourier Transform.

The melfrequency scale [17] is consider as linear frequency having spacing less than 1000 Hz and a logarithmic spacing more than 1000Hz .As a reference point ,a pitch of a 1 KHz tone ,40 dB above the threshold perceptual hearing, is defined as 1000 mels. So, to find the mels for a specific given frequency f in Hz we can use this equation 2.

$$\mathrm{Mel}(f) = 2595*\log10(1 + f/700) \qquad (2)$$

The MFCC features correspond to the total power of the log in a critical band around the center frequencies. Finally, for the calculation of cepstral coefficients, the Inverse Discrete Fourier Transformer is applied; finally calculate the DCT of the output from the filter bank. The resultant value is the actual Mel-Frequency Cepstral Coefficient.

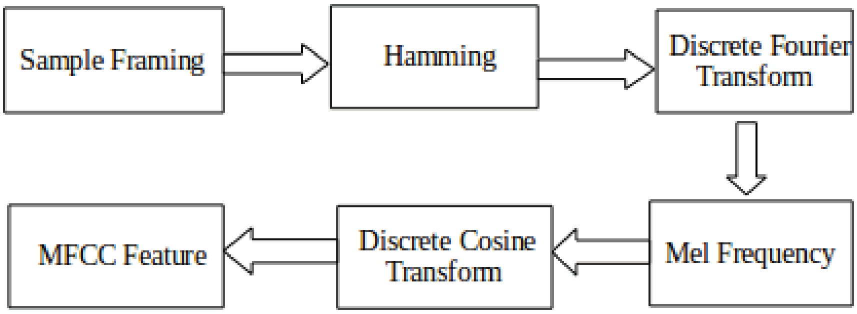


Figure 1: MFCC features Extraction Technique

B Perceptual Linear Predictive

The PLP model aims at human vocalizations based on the concept of hearing psychophysics and then more precisely in the process of extracting features. PLP increases the rate of speech recognition and also eliminates irrelevant speech information [18]. PLP technique is quite similar to LPC but differs from MFCC. PLP mainly consists of three steps. First one is for critical band analysis. Second is for equal loudness and the third one is for intensity-loudness and power-law relation. PLP carries out spectral analysis with frame of N samples with N band filters on the speech vector. For the experiments, 256 window sizes and 24 filter banks are used. The PLP filters are then produced with pre-emphasis and scale of bark. Next step is the estimation of power spectrum with the power law [18]. Now computed PLP spectrum is forwarded to LP analysis with the frequencies. At-last LP analysis is performed along FFT and then final values are observed by calculating the inverse of FFT.

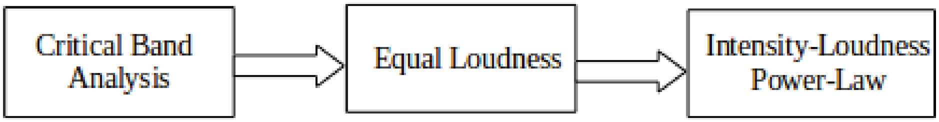


Figure 2: PLP features Extraction Technique

C Linear Prediction Coefficient

The LPC is actually works on the prediction. In samples of speech signal, we can predict the nth samples, which can be represented by summarizing the previous samples of the target signals (k). The inverse filter production should be carried out to match the formants region of the speech samples [19]. The LPC process is therefore the application of these filters in the samples [20]. The main idea of LPC is to approximate the current (n) acoustic sample s(n) with the previous samples s(p).

$$s(n) \approx a_1(n - 1) + a_2(n - 2) \pm \cdots \pm a_p(n - p) \qquad (3)$$

Then LPC is obtained using the Levinson-Durbin recursive algorithm [20].

$$H(z) = \frac{\mathbf{1}}{1 - \sum_{k-1}^{p} a_k z^{-k}} \qquad (4)$$

H(z) reflects the propagation path of the acoustic signal. Let c (n) be the impulse response [20]:

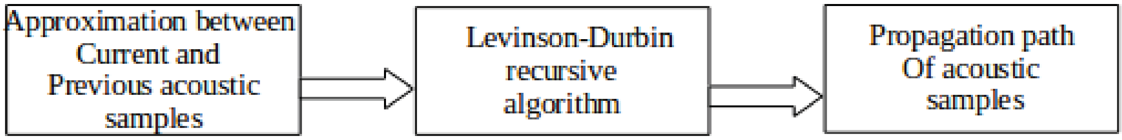

Approximation between Current and Previous acoustic samples
Levinson-Durbin recursive algorithm
Propagation path Of acoustic samples

Figure 3: LPC features Extraction Technique

$$H^{Z} = \ln H(z) = \sum_{n=0}^{00} c(n) z^{-k} \qquad (5)$$

D Linear Prediction Cepstral Coefficient

Linear Prediction Cepstral is an enhanced version of LPC method. The representation of linear predictive coefficients is in cepstrum domain can be reflected as new coefficients known as linear predictive cepstral coefficients [21]. The value of LPCC coefficient can be computed by using LPC equations which are as follows.

$$C_1 = a_1 \tag{6}$$

$$c_n = a_n + \sum_{k-1}^{n-1} \frac{k}{n} \; c_k a_{n-k} 1 < n \leq p \tag{7}$$

$$c_n = a_n + \sum_{k-1}^{n-1} \frac{k}{n} c_k a_{n-k} n > p \tag{8}$$

Where $C_1, C_2, \cdots C_n$ are the LPCC.

E Principle Component Analysis (PCA)

The PCA is thought a Principle part Analysis – this is often a statistical analytical tool that's used to explore kind and cluster information. PCA take an over-sized variety of correlate (interrelated) variables and rework this information into a smaller variety of unrelated variables (principal components) whereas holding largest quantity of variation, so creating it easier to work the information and build predictions. PCA could be a method of distinguishing patterns in information, and expressing the information in such some way on highlight their similarities and variations. Since a pattern in information is hard to seek out in information of high dimension, wherever the posh of graphical illustration isn't offered, PCA could be a powerful tool for analyzing information [10].

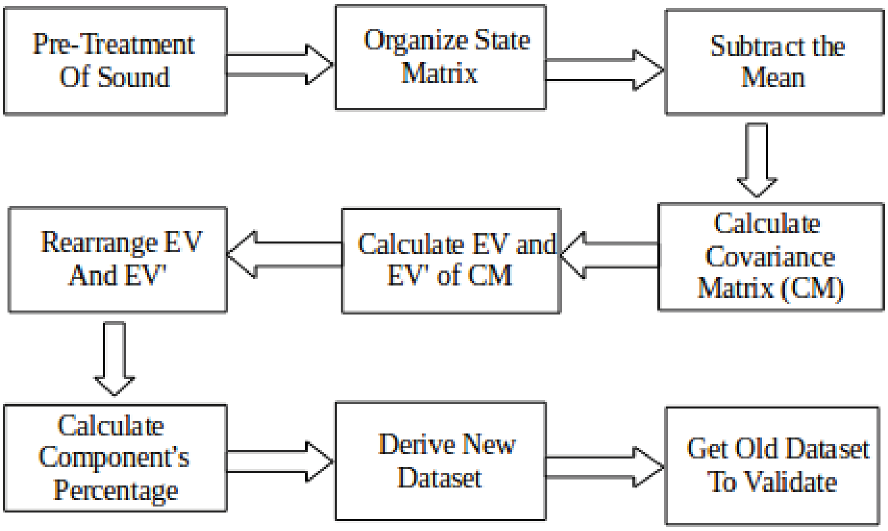

Pre-Treatment Of Sound
Organize State Matrix
Subtract the Mean
Calculate Covariance Matrix (CM)
Calculate EV and EV' of CM
Rearrange EV And EV'
Calculate Component's Percentage
Derive New Dataset
Get Old Dataset To Validate

Figure 4: PCA features Extraction Technique

Where EV is Eigen Vector and EV' is Eigen Value.

F *Relative Spectral Processing*

The RASTA is method of extracting the relevant information from a sound or any speech signal and the main objective of this technique is to eliminate the robustness of speech recognition in noise or in the real time environments [16] and it is usually done by using time trajectories of band pass filter of logarithmic speech value, infact it is the extension of the original method by combining additive noise and convolution noise [15].

RASTA is a voice improvement based on linear filtering of the short-term power spectrum of the noisy audio signal, as shown in Figure 5. The input speech signal spectral values are compressed by a nonlinear compression rule (a = 2/3) before filtering and expanded after filtering (b = 3/2) [16].

Output of each filter is given as,

$$S_i\,(\mathrm{K}) = \sum_{j=-M}^{M} W_i\,(j) Y_i\,(k) \qquad (9)$$

Si(k) is a clean speech estimate and Yi(k) is the noisy audio spectrum, Wi(j) is the filter weights and M is the filter order.

These values can be set to the required processing or the corresponding set of problem.

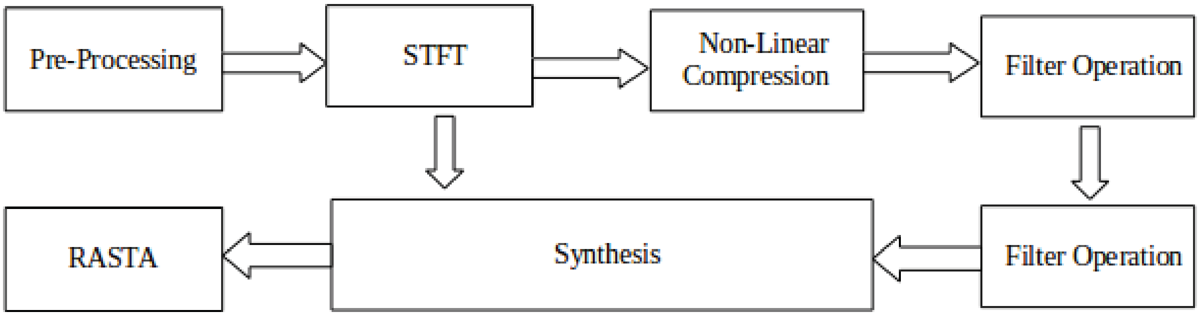


Figure 5: RASTA features Extraction Technique

4 Results

Details comparison observed during this survey are following that will help researchers and practitioner to know the insights of different feature extraction techniques used in sound recognition and detection problems.

Table 1: Comparative Analysis of Features Extraction Techniques Used in Sound Recognition and Detection

Features Extraction	Noisy Data	Complexity	Accuracy
MFCC	Poor result on noisy data [27]	Less Complex and High performance rate	92% [24]
PLP	poor result on noisy data due to spectral balance of formant	Slightly Complex	Better Performance than LPCC and MFCC [25]
PCA	Doesn't work well on noisy data as it does not reduce noise completely.	Slightly Complex and High Performance Rate	54.66% [8]
LPC	Not good for noisy data [27].	Less complex [27].	Good Accuracy, reliability and robustness [24]
LPCC	Shows poor result on highly noised data [27].	Simple and good performance [27]	Accuracy is 88% [26]
RASTA	Works good on noisy data as it enhances data by removing noisy distortions [27].	Slightly Complex	A robust technique. Low modulation frequencies are captured [24].

Table 2: Comparative Analysis of Features Extraction Techniques Used in Sound

Features Extraction	Extraction Method	Final Comments
MFCC	Dynamic method [27].	Mostly used where human ear bandwidth scale exists.
PLP	Combines the linear prediction analysis and spectral analysis [9]	Increases the recognition rate and also removes noise.
PCA	Non-Linear method [27].	Eigen vector based. Reduce Components / Dimensions of Features
LPC	A static method [29].	Used for extraction at lower rate. It can be used in sound recognition of abnormal sounds
LPCC	Use Autocorrelation analysis [27].	Used in cepstral domain.
RASTA	Non-Linear Compression [16].	Highly recommended in domain where there is noise, it will extract good features in noisy data

5 Conclusion

In this paper we have discussed some widely used feature extraction techniques in the domain of speech recognition. The motivation for doing this comparative analysis is because there are many feature extraction techniques that are available and very few of them are really helpful. This paper will guide the researchers for methods feature extraction technique and it will also help them to differentiate between different Novel and Robust features can also be extracted by combining many of the existing features to enhance the capability of sound detection and recognition systems. Developing a system that will record complete meeting conversation in a dialogue form, sentence spoken by each person against their name (if known), otherwise a separate line by some person “i”. This system will reduce time of recording meeting or writing manual points where some points may skipped or interpreted wrong.

References

[1] McKinney, Martin, and Jeroen Breebaart. "Features for audio and music classification." (2003).

[2] Davis, Steven, and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences."IEEE transactions on acoustics, speech, and signal processing 28, no. 4 (1980): 357-366.

[3] Wold, Erling, Thom Blum, Douglas Keislar, and James Wheaten. "Content-based classification, search, and retrieval of audio."IEEE multimedia 3, no. 3 (1996): 27-36.

[4] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In Proc. ICASSP, pages 1331–1334, Munich, Germany, 1997.

[5] H. Hermansky and N. Malayath. Spectral basis functions from discriminant analysis. In International Conference on Spoken Language Processing, 1998

[6] M. Zhang, K. Tan, and M. H. Er. Three-dimensional sound synthesis based on head-related transfer functions. J. Audio. Eng. Soc., 146:836–844, 1998.

[7] T. Zhang and C. C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on speech and audio processing, 2001.

[8] Prasad, K.S., Ramaiah, G.K. and Manjunatha, M.B., 2017. Speech features extraction techniques for robust emotional speech analysis/recognition. Indian Journal of Science and Technology.

[9] Vimala, C. and Radha, V., 2014. Suitable feature extraction and speech recognition technique for isolated tamil spoken words. International Journal of Computer Science and Information Technologies (IJCSIT), 5(1), pp.378-383.

[10] Shrawankar, U. and Thakare, V.M., 2013. Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:1305.1145.

[11] Dave, N., 2013. Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6), pp.1-4.

[12] Barchiesi, D., Giannoulis, D., Stowell, D. and Plumbley, M.D., 2015. Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 32(3), pp.16-34.

[13] Desai, N., Dhameliya, K. and Desai, V., 2013. Feature extraction and classification techniques for speech recognition: A review. International Journal of Emerging Technology and Advanced Engineering, 3(12), pp.367-371.

[14] Këpuska, V.Z. and Elharati, H.A., 2015. Robust speech recognition system using conventional and hybrid features of mfcc, lpcc, plp, rasta-plp and hidden markov model classifier in noisy conditions. Journal of Computer and Communications, 3(06), p.1.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, Oct. 1994.

[16] Kurzekar, P.K., Deshmukh, R.R., Waghmare, V.B. and Shrishrimal, P.P., 2014. A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology, 3(12), pp.18006-18016.

[17] Tiwari, V., 2010. MFCC and its applications in speaker recognition. International journal on emerging technologies, 1(1), pp.19-22.

[18] Chandra, E., and K. Manikandan M. Sivasankar. "A proportional study on feature extraction method in automatic speech recognition system." (2014).

[19] Narang, S. and Gupta, M.D., 2015. Speech Feature Extraction Techniques: A Review. International Journal of Computer Science and Mobile Computing, 4(3), pp.107-114.

[20] Yang, S., Cao, J. and Wang, J., 2015, July. Acoustics recognition of construction equipments based on LPCC features and SVM. In Control Conference (CCC), 2015 34th Chinese (pp. 3987-3991). IEEE.

[21] Kaur, K. and Jain, N., 2015. Feature Extraction and Classification for Automatic Speaker Recognition System-A Review. International Journal of Advanced Research in Computer Science and Software Engineering, 5.

[22] Gupta, D., Bansal, P. and Choudhary, K., 2018. The state of the art of feature extraction techniques in speech recognition. In Speech and Language Processing for Human-Machine Communications (pp. 195-207). Springer, Singapore.

[23] Chaudhary, G., Srivastava, S. and Bhardwaj, S., 2017. Feature Extraction Methods for Speaker Recognition: A Review. International Journal of Pattern Recognition and Artificial Intelligence, 31(12), p.1750041.

[24] Gupta, H. and Gupta, D., 2016, January. LPC and LPCC method of feature extraction in Speech Recognition System. In Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference (pp. 498-502). IEEE.

[25] Khara, S., Singh, S. and Vir, D., 2018, April. A Comparative Study of the Techniques for Feature Extraction and Classification in Stuttering. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 887-893). IEEE.

[26] Kianisarkaleh, A. and Ghassemian, H., 2016. Spatial-spectral locality preserving projection for hyperspectral image classification with limited training samples. International journal of remote sensing, 37(21), pp.5045-5059.

[27] Singh, P.P. and Rani, P., 2014. An approach to extract feature using mfcc. IOSR Journal of Engineering, 4(8), pp.21-25.

[28] Kaur, J. and Sharma, A., 2014. Emotion detection independent of user using mfcc feature extraction. International Journal of Advanced Research in Computer Science and Software Engineering, 4(6).

[29] Dhingra, S.D., Nijhawan, G. and Pandit, P., 2013. Isolated speech recognition using MFCC and DTW. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(8), pp.4085-4092.

[30] Poonkuzhali, C., Karthiprakash, R., Valarmathy, S. and Kalamani, M., 2013. An approach to feature selection algorithm based on ant colony optimization for automatic speech recognition. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(11), pp.5671-5678.

[31] Kuncheva, L.I. and Faithfull, W.J., 2014. PCA feature extraction for change detection in multidimensional unlabeled data. IEEE transactions on neural networks and learning systems, 25(1), pp.69-80.

[32] Zabalza, J., Ren, J., Yang, M., Zhang, Y., Wang, J., Marshall, S. and Han, J., 2014. Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 93, pp.112-122.

[33] Ren, J., Zabalza, J., Marshall, S. and Zheng, J., 2014. Effective feature extraction and data reduction in remote sensing using hyperspectral imaging [applications corner]. IEEE Signal Processing Magazine, 31(4), pp.149-154.

[34] Hibare, R. and Vibhute, A., 2014. Feature extraction techniques in speech processing: a survey. International Journal of Computer Applications, 107(5).

[35] Zouhir, Y. and Ouni, K., 2014. A bio-inspired feature extraction for robust speech recognition. SpringerPlus, 3(1), p.651.

[36] Gulzar, T., Singh, A. and Sharma, S., 2014. Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks. International Journal of Computer Applications, 101(12), pp.22-27.

[37] Hermansky, H. and Fousek, P., 2005. Multi-resolution RASTA filtering for TANDEM-based ASR (No. REP_WORK). IDIAP.

Improving Requirement Prioritization process in Product line using Artificial Intelligence technique

Wasi Haider[1] Yaser Hafeez[2] Sadia Ali[3]

M Azeem Abbas[4] M Numan Rafi[5] Abdul Salam[6]

Abstract

Product families emerged a new and useful development technique in the field of software development. In Software Product Line (SPL) there are some core assets and some variants so using these assets anyone can build their desired product in very short time and effort. While working in product family's environment we must keep an eye on the requirement prioritization and ranking because that requirement are very important because these requirement lay the foundation of the core and variants assets which are the building blocks of SPL. So there are some major issues which we face are the more human interaction, ambiguous requirements and wrong or no requirement ranking. In this paper we proposed a framework for the ranking of stakeholders' requirements for the SPL's variant and core assets using the case base reasoning CBR if available in previous use or assign them new ranking according to their requirement and their assign ranking for software product line. We evaluated our framework by empirical study. The results prove that the considerable improvement for different parameters is achieved by our framework as compared to conventional approaches of requirement prioritization.

Keyword: Software Product Line (SPL); Requirement Prioritization (RP); Case Base Reasoning (CBR); Artificial intelligence (AI)

1 Introduction

Software product family is a interrelated software systems, sharing a common and managed collection of features to accomplish the wants of a suitable market section [1]. The main goal of SPL is reuse in an effort to enhance the quality and production while reducing cost as well as time to market. SPL engineering has become an efficient and minimizes time-to-develop approach for providing a common model for developing product families. The central concept at the back of SPL is to provide a stage with common and distinct components of a software system identified in order to build a consistent line of products [2]. Software product variants are often develop from an early product development. These product variants are generally share some common but they are also different from each other due to upcoming change request to fulfill the specific demand and requirement of the end user [3]. As a number of features and the number of products increase, it is significance re-engineering product variants into a SPL

[1] PMAS-Arid Agriculture University, Rawalpindi | wasihaider734@gmail.com
[2] PMAS-Arid Agriculture University, Rawalpindi | yasir@uaar.edu.pk
[3] PMAS-Arid Agriculture University, Rawalpindi | sadiaalief@gmail.com
[4] PMAS-Arid Agriculture University, Rawalpindi | azeem.abbas@uaar.edu.pk
[5] PMAS-Arid Agriculture University, Rawalpindi | miannuman10437@gmail.com
[6] PMAS-Arid Agriculture University, Rawalpindi | abdulsalam2301@gmail.com

for systematic reuse. The first step of SPLE is to extract a feature model. This further suggested recognizing the common and variant features. Manual reverse engineering of feature model for the available software variants is time and effort taking [4].

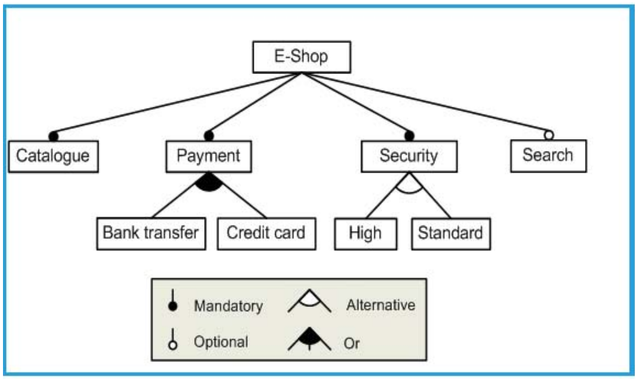


Figure 1: SPL Feature Model

When developing software, Requirements Engineering is field of defining, documenting and maintaining software requirements, mostly described in natural language [5]. This information motivated some proposals to use Natural Language Processing (NLP) to minimize the uncertainty, identify omitted information, and even enhance traceability with remaining stage of process [6].

Requirement prioritization (RP) is a main part in the requirement engineering phase. RP plays a vital part in the RE process, particularly, regarding vital tasks like requirements negotiation and software release [7]. Outstanding RP is necessary to any well-run project. It ensures that project concentrate on the main parts first, and that everybody perceived and conforms about what the project's most important parts are. There are many techniques, which are helpful for specification and prioritization of requirement according to stakeholder's time, cost, nature of the project etc. When developer used any requirement prioritization technique and find out the priority or ranking of requirements in any system, they save the ranking of the requirement with it all information and stored it in database for reuse purpose in future. For knowledge management and reuse of previous knowledge, researcher adapted AI technique called case based reasoning (CBR). CBR retrieve previous solutions for current problem solving base on expert knowledge intelligently in different scenarios [8].

In this paper, we have presented a comprehensive framework for the requirements ranking in which we extract the commonalities Cs and variabilities Vs of the software product line from the requirement document using J48 Decision algorithm. It initiates the rules for the calculation of the target variable. With the assistance of J48 classification algorithm [9] the significant distribution of the data is easily understandable. After finding the Cs and Vs apply the CBR and find the previous ranking if available then assign them else assign their ranking and find out the sorted prioritized requirement list.

The rest of this paper is structured as follows: Section 2 present literature review. In Section 3, we present our framework. In Section 4, we present evaluation and discussion.

2 Literature Review

The growing complication and cost of software-intensive systems has led developers to find the alternatives of reusing software parts in development of systems. One approach to increasing re-usability is to develop a SPL. Existing research has paying attention on techniques that create a configuration of an SPL in a single step. First, they present a formal model of multi-step SPL. Second, present the solutions to these SPL configuration problems can be automatically derived with a constraint. In future work, they plan to investigate Real-time configuration process monitoring [10].

The analysis of the requirements artifacts (SRS document, use case models) is a time taking process when performed manually. There is also required for creating consistent and complete collection of NFRs from user-specific individual projects in SPL. Therefore, they [11] propose a method to create Domain NFRs from Product NFRs using model driven approach.

It is essential for an organization to boost value creation for a given investment. The principle RE activities are to add business value that is considered for in terms of return on investment of a software product. This [12] paper provides insight into the release planning processes used in the software industry to create software product value. It presents to what degree the significant stakeholders' viewpoints are spoken to in the basic decision-making process.

SPLE strengthened high-level constructive software reuse by exploiting commonality and managing variability in a product family. To overcome the complexity of the modeling, it is divided into two views a feature tree and a dependency view [13].

In the development of a SPL, any project requires to grow core assets according to the change in environment, market, and technology. In order to successfully grow core assets, it is critical for the project to get ready and use a standardized strategy for prioritizing requirements. In paper [14], authors examine the evolution of foundation assets. Tacit knowledge for prioritizing requirements was extracted. Such knowledge was made explicit and clear to develop a way for prioritizing.

Reusing of software varies from operational, ad-hoc and short-term to strategic, planned and long-term. They [15] present and compare two different requirements-led approaches. The first deals with requirements reuse and re-usability in context of product line engineering and second in context of CBR. To assist large-scale development they proposed a Feature-Similarity model.

Requirements assurance seeks to maximize confidence in the quality of requirements through audit and review. Authors of [16] present a method that applies well-established text-mining and statistical methods to minimize this effort and increase traceability matrix assurance. The method is new, that it utilizes both requirements similarity and dissimilarity.

Prioritizing requirements focus on stakeholders' feedback brings a noteworthy cost because of time elapsed in a large number of human interactions. A Semi-Automated Framework has been presented in paper [17]. It predicts appropriate stakeholders' ratings to reduce human interactions. Future work of this research is to cluster requirements.

A prioritization method called Case-Based Ranking (CB Rank), presented in [18] which integrate project's stakeholder's desires with requirements ordering approximations calculated through AI techniques.

3 Methodology

In this segment, we present our proposed framework for ranking of stakeholders' requirements using the case base reasoning CBR if available in previous use or assign them new ranking according to their requirement and their assign ranking for software product line.

A Proposed Approach

Our framework gives an inclusive model for the requirement ranking of software product line using the CBR. Our proposed framework consists of the following layers which are:

1) Description Layer:

In this first layer we performed profiling of the system, it include two main steps first is requirement elicitation which is the process of extracting the information from stakeholders. We also get the initial ranking from the stakeholders against each requirement. As the outcome of this layer we get the requirement document along with requirement initial ranking.

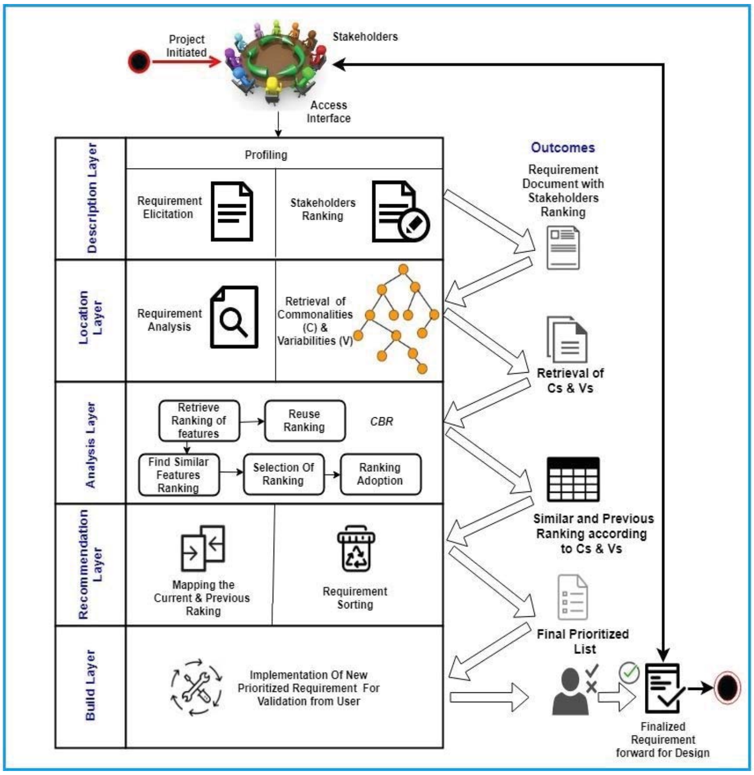

Project Initiated
Stakeholders
Access Interface
Description Layer
Profiling
Requirement Elicitation
Stakeholders Ranking
Location Layer
Requirement Analysis
Retrieval of Commonalities (C) & Variabilities (V)
Analysis Layer
Retrieve Ranking of features
Reuse Ranking
CBR
Find Similar Features Ranking
Selection Of Ranking
Ranking Adoption
Recommendation Layer
Mapping the Current & Previous Raking
Requirement Sorting
Build Layer
Implementation Of New Prioritized Requirement For Validation from User
Outcomes
Requirement Document with Stakeholders Ranking
Retrieval of Cs & Vs
Similar and Previous Ranking according to Cs & Vs
Final Prioritized List
Finalized Requirement forward for Design

Figure 2: Proposed Framework

2) Location Layer:

In this second layer, we find the commonalities and variabilities of product line from the document using the J48 classification algorithm. It generates a binary tree. This approach is helpful in classification problem. Using this technique, a tree is constructed to model the classification process. [19]

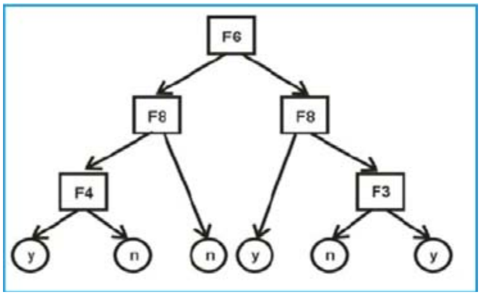


Figure 3: J48 Working

3) Analysis Layer:

In analysis Layer we apply the CBR, it is an AI technique that work on expert knowledge and previous experiences with less time, effort and cost. It works on the concept of reuse the solution of previous cases like new case and stores the cases in the database for later use. [8]

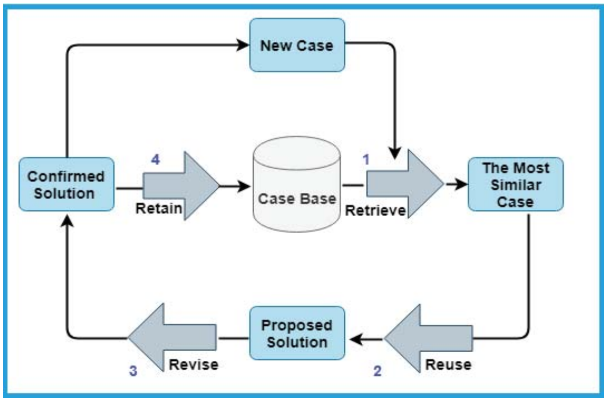


Figure 4: Case-Based Ranking (CBR)

4) Recommendation Layer:

In this layer we map the ranking of the stakeholder's requirements and the ranking find out from the CBR if we found the better result against the applied query we adopt the best available ranking and then apply the sorting on that list and we get the sorted prioritized list as the outcome of this layer.

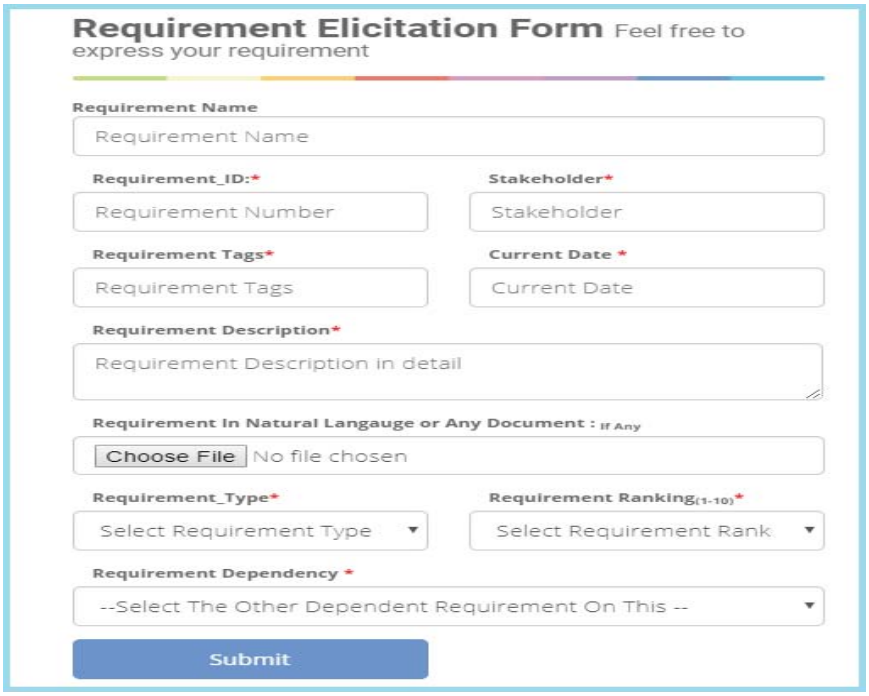
5) Build Layer:

At this last layer we send the prioritized list to the stakeholders if they accept it and approved it then we forward it to the prototype and design of the product

4 Results and Discussion

For practical implementation of our proposed work in real world context we developed an intelligent requirements prioritization recommendation (IRPR) tool using steps of proposed work. Therefore, to evaluate IRPR we performed an empirical study. For this matter of fact, we technologies development organization which work on different projects both nationally and globally, but company not allow us to disclose any information about company. From large bulk of projects pool we selected two projects (P) i.e. LMS system (P-A) and card swipe machine (P-B).

For the elicitation and prioritization of projects user requirements before implementations uses different applications. Hence, the traditional tools/techniques (TT) they adopted increase the challenges that mention in literature review section i.e. more human interaction, ambiguous requirements etc. To resolve these issues company agreed to use IRPR tool to attain higher user satisfaction and product quality. Consequently, for IRPR implementation we conducted experiment and divided participants of company employees i.e. 21 in total for experiment into two groups' i.e. experimental treatment (ET) and non-experimental treatment (NET). The participants of ET used to develop both P-A and P-B using IRPR whereas NET participants adopted TT for implementation both projects. While participants consist of project manager (PM), requirement engineers (RE), requirement analysis (RA), developers (D) and stakeholders (S). The overall working of IRPR prototype show in figure 5-9 to illustrate the interfaces of IRPR.



Requirement Elicitation Form Feel free to express your requirement

Requirement Name
Requirement Name

Requirement_ID:*
Requirement Number

Stakeholder*
Stakeholder

Requirement Tags*
Requirement Tags

Current Date *
Current Date

Requirement Description*
Requirement Description in detail

Requirement In Natural Langauge or Any Document : If Any
Choose File No file chosen

Requirement_Type*
Select Requirement Type ▾

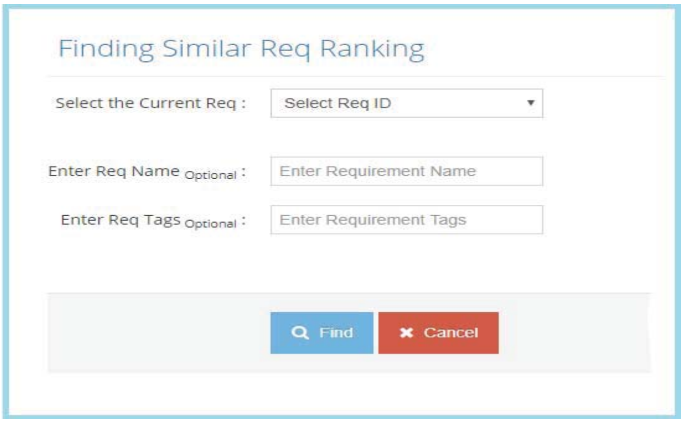
Requirement Ranking(1-10)*
Select Requirement Rank ▾

Requirement Dependency *
--Select The Other Dependent Requirement On This -- ▾

Submit

Figure 5: Requirement Elicitation & Stakeholders Ranking

When the project initiated the working of IRPR started; therefore, S connected to PM and the form open for elicitation of requirements as show in figure 5 screen shot of form interface. In the form user enter their requirements with ranking and profile of all users maintained in the database for future use. After the evaluation of profiling RE and RA analysis the requirements because these projects are SPL based. Therefore, then using j48 algorithm retrieve commonalities and variabilities in the form of decision tree for the CBR mapping to extract previous ranking as depicted in figure 6.



Finding Similar Req Ranking

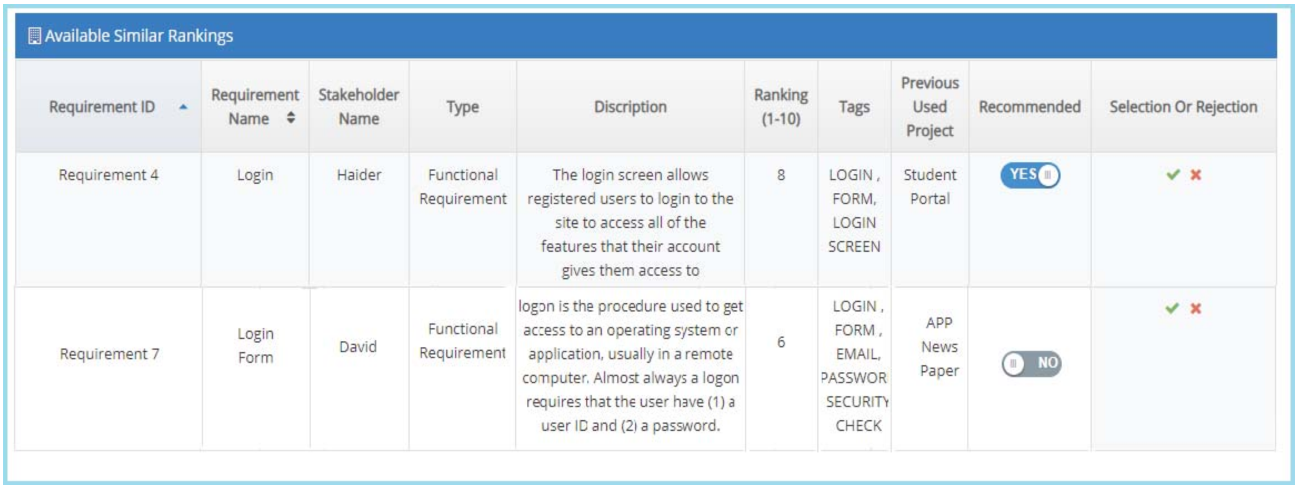
Select the Current Req : Select Req ID ▾

Enter Req Name Optional : Enter Requirement Name

Enter Req Tags Optional : Enter Requirement Tags

Figure 6: Finding Similar Ranking Query (CBR)

In CBR when we apply a query for finding similar ranking we will get the list (shown in figure 7) of the previous cases which are similar to the current case with the ranking. We will accept and adopt the case which is high rank amongst them.

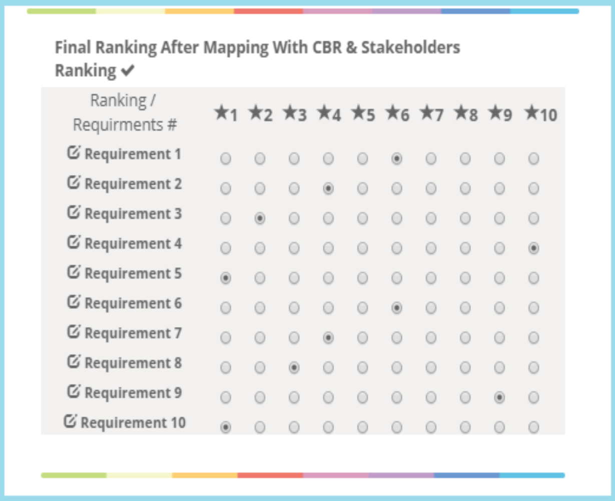


Available Similar Rankings

Requirement ID	Requirement Name	Stakeholder Name	Type	Discription	Ranking (1-10)	Tags	Previous Used Project	Recommended	Selection Or Rejection
Requirement 4	Login	Haider	Functional Requirement	The login screen allows registered users to login to the site to access all of the features that their account gives them access to	8	LOGIN , FORM, LOGIN SCREEN	Student Portal	YES	✓ ✗
Requirement 7	Login Form	David	Functional Requirement	logon is the procedure used to get access to an operating system or application, usually in a remote computer. Almost always a logon requires that the user have (1) a user ID and (2) a password.	6	LOGIN , FORM , EMAIL, PASSWOR SECURITY CHECK	APP News Paper	NO	✓ ✗

Figure 7: Selection and adoption of similar ranking

When we adopt some cases from CBR and mapped the current and the previous ranking we will get the prioritized list of the requirement with the ranking from 1-10 in a unsorted order shown in Figure 8 below.

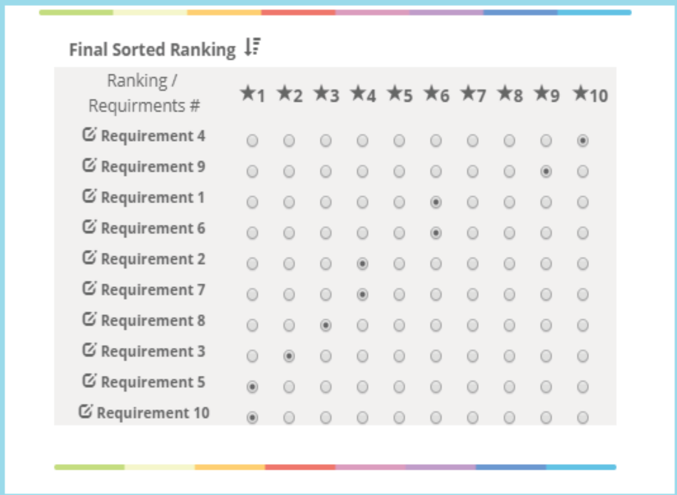


Final Ranking After Mapping With CBR & Stakeholders Ranking ✓

Ranking / Requirments #	★1	★2	★3	★4	★5	★6	★7	★8	★9	★10
Requirement 1	○	○	○	○	○	●	○	○	○	○
Requirement 2	○	○	○	●	○	○	○	○	○	○
Requirement 3	○	●	○	○	○	○	○	○	○	○
Requirement 4	○	○	○	○	○	○	○	○	○	●
Requirement 5	●	○	○	○	○	○	○	○	○	○
Requirement 6	○	○	○	○	○	●	○	○	○	○
Requirement 7	○	○	○	●	○	○	○	○	○	○
Requirement 8	○	○	●	○	○	○	○	○	○	○
Requirement 9	○	○	○	○	○	○	○	○	●	○
Requirement 10	●	○	○	○	○	○	○	○	○	○

Figure 8: Final Prioritized Requirements after mapping stakeholders and CBR Ranking

Apply any sorting technique with respect to their ranking we will get the final sorted prioritized ranking of the requirements (shown in figure 9) which will decide the education order of the requirement in development phase



Final Sorted Ranking ↓=

Ranking / Requirments #	★1	★2	★3	★4	★5	★6	★7	★8	★9	★10
Requirement 4	○	○	○	○	○	○	○	○	○	●
Requirement 9	○	○	○	○	○	○	○	○	●	○
Requirement 1	○	○	○	○	○	●	○	○	○	○
Requirement 6	○	○	○	○	○	●	○	○	○	○
Requirement 2	○	○	○	●	○	○	○	○	○	○
Requirement 7	○	○	○	●	○	○	○	○	○	○
Requirement 8	○	○	●	○	○	○	○	○	○	○
Requirement 3	○	●	○	○	○	○	○	○	○	○
Requirement 5	●	○	○	○	○	○	○	○	○	○
Requirement 10	●	○	○	○	○	○	○	○	○	○

Figure 9: Sorted Prioritized Requirements

For the assessment of experiment performance, we conducted questioner based review from both ET and NET members. The review based on parametric analysis which based on existing literate i.e. user friendly (UF), usability (U), learnability (L), efficient (E), high effectiveness (HE), less human interaction (LHI), proficient knowledge management (PKM), efficient knowledge identification and retrieval (EKIR), requirements priority accuracy (RPA), enhance elicitation and prioritization (EEP), high productivity (HP) and higher user satisfaction (HUS).

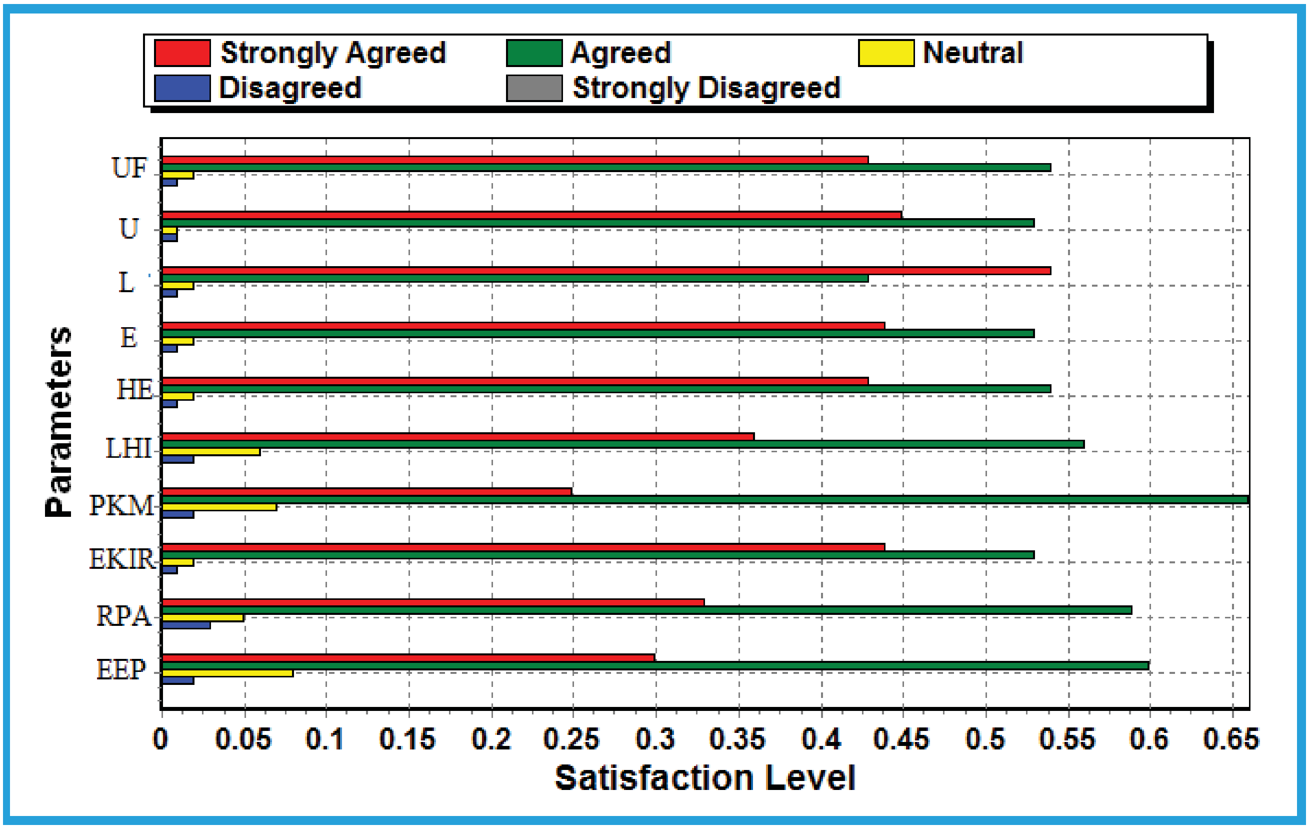


Figure 10: Review Analysis

The overall analysis result on the parameters implementing both tools i.e. IRPR and TT as demonstrate in figure 10. The figure 10 shows the satisfaction ratio of users on left side vertically with more than 50 percent satisfaction ratio and parameters review on the Y-axis.

Table 1: Comparative Analysis

Techniques	Participators				
	PM	TL	RA	Ds	QE
Experimental Treatment of P-A (ET P-A)	0.7	0.69	0.63	0.86	0.76
Non- Experimental Treatment of P-A (NET P-A)	0.32	0.36	0.45	0.27	0.38
Experimental Treatment of P-B (ET P-B)	0.80	0.70	0.60	0.76	0.86
Non-Experimental Treatment of P-A (NET P-B)	0.35	0.40	0.45	0.37	0.28

The users of project A in which experimental treatment (ET) is applied, are more satisfied and gained better results than the participants of non-experimental treatment (NET). Whereas; same is the case with participants of project B. The members of experimental treatment (ET) of B give better quality and competence.

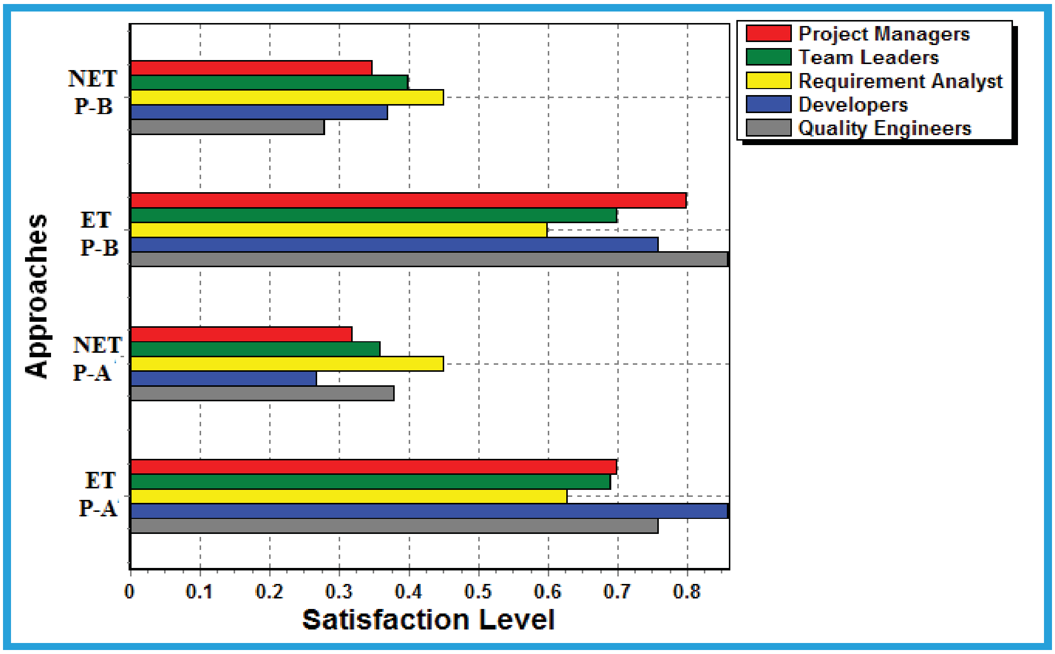


Figure 11: Comparative analysis results

We also illustrate the comparative analysis of both projects with experimental treatment (ET) and non-experimental treatment (NET) in Figure 11. Figure 11 represent the participants' satisfaction level. The y-axis labels each project development approaches while x-axis explains the satisfaction level of each user. The results present that our proposed framework's performance and satisfaction for quality and customer needs.

5 CONCLUSION

In this research, we proposed a framework for requirement ranking for software product line using CBR. The proposed framework uses J48 to find out the Cs and Vs from requirement document and then apply CBR on these requirements to find their final ranking. We have performed a tool based evaluation to evaluate our framework. Our results show noteworthy improvement in terms of satisfaction level for various parameters as compared to traditional approaches of ranking in SPL. The proposed research provides direction to industry and researchers to manage software prioritization.

References

[1] Bhushan, Megha, Shivani Goel, and Karamjit Kaur. "Analyzing inconsistencies in software product lines using an ontological rule-based approach." Journal of Systems and Software 137 (2018): 605-617.

[2] Khalique, F., Butt, W.H. and Khan, S.A., 2017, December. Creating domain non-functional requirements software product line engineering using model transformations. In 2017 International Conference on Frontiers of Information Technology (FIT) (pp. 41-45). IEEE.

[3] Xue, Y., Xing, Z., Jarzabek, S.: Feature location in a collection of product variants. In: IEEE 19th RE Conference, pp. 145–154 (2012)

[4] Ra'Fat, A., Seriai, A., Huchard, M., Urtado, C., Vauttier, S. and Salman, H.E., 2013, June. Feature location in a collection of software product variants using formal concept analysis. In International Conference on Software Reuse (pp. 302-307). Springer, Berlin, Heidelberg.

[5] D. Zowghi and C. Coulin, Requirements Elicitation: A Survey of Techniques, Approaches, and Tools, pp. 19–46. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[6] Arias, M., Buccella, A. and Cechich, A., 2018. A Framework for Managing Requirements of Software Product Lines. Electronic Notes in Theoretical Computer Science, 339, pp.5-20.

[7] Hasan, M. S., Mahmood, A. Al, Alam, J., & Hasan, S. N. (2010). An Evaluation of Software Requirement Prioritization Techniques. International Journal of Computer Scienceand Information Security, 8(9), 83–94.

[8] Ali, S., Iqbal, N. and Hafeez, Y., 2018. Towards Requirement Change Management for Global Software Development using Case Base Reasoning. Mehran University Research Journal of Engineering and Technology, 37(3), pp.639-652.

[9] Kaur, G. and Chhabra, A., 2014. Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22).

[10] White, J., Galindo, J.A., Saxena, T., Dougherty, B., Benavides, D. and Schmidt, D.C., 2014. Evolving feature model configurations in software product lines. Journal of Systems and Software, 87, pp.119-136.

[11] Khalique, F., Butt, W.H. and Khan, S.A., 2017, December. Creating domain non-functional requirements software product line engineering using model transformations. In 2017 International Conference on Frontiers of Information Technology (FIT) (pp. 41-45). IEEE.

[12] Barney, S., Aurum, A. and Wohlin, C., 2008. A product management challenge: Creating software product value through requirements selection. Journal of Systems Architecture, 54(6), pp.576-593.

[13] Ye, H. and Liu, H., 2005. Approach to modelling feature variability and dependencies in software product lines. IEE Proceedings-Software, 152(3), pp.101-109.

[14] Inoki, M., Kitagawa, T. and Honiden, S., 2014, August. Application of requirements prioritization decision rules in software product line evolution. In Requirements Prioritization and Communication (RePriCo), 2014 IEEE 5th International Workshop on (pp. 1-10). IEEE.

[15] Kaindl, H. and Mannion, M., 2018, August. Software Reuse and Reusability based on Requirements: Product Lines, Cases and Feature-Similarity Models. In 2018 IEEE 26th International Requirements Engineering Conference (RE) (pp. 510-511). IEEE.

[16] Port, D., Nikora, A., Hayes, J. H., & Huang, L. (2011, January). Text mining support for software requirements: Traceability assurance. In System Sciences (HICSS), 2011 44th

Hawaii International Conference on (pp. 1-11). IEEE.

[17] Asif, S. A., Masud, Z., Easmin, R., & Ul, A. (n.d.). arXiv : 1801 . 00354v1 [cs . SE] 31 Dec 2017 SAFFRON : A Semi-Automated Framework for Software Requirements Prioritization, 1–21

[18] Perini, A., Susi, A., & Avesani, P. (2013). A Machine Learning Approach to Software Requirements Prioritization, 39(4), 445–461.

[19] Patil, T.R. and Sherekar, S.S., 2013. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), pp.256-

Real-Time Simulation of IoT Based Smart Home System and Services Using RFID

Rana M. Amir Latif[1] Muhammad Farhan[2] Laiqa Binte Imran[3]

Kashif Manzoor[4] Tayyaba Tariq[5] Hassan Raza[6]

Abstract

IoT plays a remarkable role in human life, By using the services of IoT, workload and pressure on human life is decreasing. As the use of IoT in smart homes and smart societies can change the lifestyle of humans, it can also play a vital role to save human life. Human life is a significant factor that is always ignored by human itself. Anything in this world cannot replace the life of any human. In this paper, we are going to facilitate our users with the smart home. For example, the tap water remains open and little children play freely in home so there could be a chance of drawing up of that child but using IoT we can inform the users that their tap is open, when the Gas cylinder leaks then IoT can also inform users about the leakage of gas and so on. We simulate the IoT based architecture diagram in Porteous simulator. These new IoT concepts in smart homes can be applied along with RFID technologies. This technology provides us with benefits in terms of cost to save human life, energy consumption, and complexity. Many smart home use cases such as water taps, washing machine, kitchen, gas cylinder,and so on, are described in examples that make use of this system. The RFID reader reads the information from home devices and sends the information to the Android device so that users can play more smartly in homes and human efforts reduce more precisely.

Keyword: Smart Civilization, Smart Homes, Internet of Things & RFID, GSM, Arduino UNO

1 Introduction

Smart home technology utilizes apparatus linked for the Internet of Thing (IoT) to automate and track home techniques. It stands out for self-monitoring identification and reporting technologies. The technology was initially created by IBM and has been known for predictive collapse investigation. The very first modern smart home technology goods became open to shoppers amongst 1998 along with the Ancient 2000s. Smart house technology permits users to command and track their connected home apparatus from wise residence programs, smart-phones, or alternative networked apparatus. Users may control linked house processes if they are dwelling or are even away. That enables more effective electricity and electrical usage as nicely as ensuring that the house is not secure. Smart house technology leads to wellness along with well-being enhancement by adapting individuals who have unique wants, especially elderly Men and women. The smart house technology is presently used to create wise metropolitan

[1] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | ranaamir1061@gmail.com*
[2] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | farhansajid@gmail.com*
[3] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | laiqa.imran50@gmail.com*
[4] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | kashifmanzoor028@gmail.com*
[5] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | tayyaba.tariq.tt@gmail.com*
[6] *COMSATS University Islamabad, Sahiwal Campus, Sahiwal | mr.hassanraza@outlook.com*

areas. An intelligent city acts such as an intelligent residence, precisely where systems have been tracked to a lot more efficiently operate on the metropolitan areas and spend less [1].

Some devices at the lower end of the proposed capabilities use the existent smart home systems [2]. Typically, these devices can respond to the user's command by using a computer, tablets, and smartphones that store data by using Wi-Fi or Bluetooth. IoT carries a new method of home management, intelligence, surveillance, integration of devices also the connectivity of these devices. In these devices, intelligence may be embedded completely, this could be within a platform where all devices connect to perform some functions, or it can be combined in the platform by using the cloud. Therefore, devices in the smart home can provide enhanced benefits and include the capabilities inherent in the IoT [3]. These devices can be a static object such as lights, smart plugs and a sensor that can measure the status and physical condition of the objects, in these sensors we have actuators that can perform several operations for example; turning on/off the appliance and opening doors. Both services can be combined using several devices [4]. Also, Data and information on smart home devices can be integrated with external data and information from other IoT systems. In the example, it shows the value-added services like the health care system [5].

RFID Systems with TWN4 is an RFID/NFC reader unit supporting all the conventional RFID technologies for 13.56 MHz 134.2 KHz, and 125 KHz frequencies are packed in a 31x17.8x2.5mm footprint. It provides easy placement on the main board with the component mounted to one side. In this latest version, all pins are already soldered that will support easy implementation and quicker working. This RFID reader is more suitable for working in industrial and mobile applications due to the small size, and with the verity of external antenna options also it supports expanded temperature range. For communication with smart phones, this RFID reader and NFC is making it a healthy choice for applications [6]. From the 13.56MHz RFID platform, the reader needs to boot clock, electricity, and info for the RFID label. Power distribution into the length of the label of blanking sign periods is simply 2-3ua s. Thus, standers utilize compacted information PPM, NRZ, and altered miller. The transmitter blended provider indicates having compressed electronic information. The blended signal will be routed into the label, and the affiliated RFID chip reacts with all the requested advice, like an identification quantity or date. It was not until coming mixed signal which the label is still working out. For increased efficacy and low energy intake; info move out of RFID label into this reader employs load modulation together with all the sub-carrier [7]. Even the sub-carrier frequency is significantly different by the criteria. By employing different coded information, the modulation used by sub-carrier is going to be worked. This sub-carrier is going to be gotten by a binary branch of this company frequency. To modify the loading immunity, adjuster controlled using the sub-carrier sign can be utilized. After creating the frequency spectrum created the main benefit of modulation using sub-carrier gets evident [8].

Arduino is an open-source, hardware and applications corporation, user and project network which models and produces single-board micro-controllers and micro-controller fittings to digital construction apparatus and interactive items which may control and sense both digitally and physically. Arduino boards can be bought commercially from a pre-assembled sort or as homemade fittings [9]. Arduino board layouts work with an assortment of microprocessors and controls.

The boards have been built with collections of analog and digital input/output (I/O) hooks which could be interfaced to several enlargement boards or breadboards along with also other circuits. The boards comprise sequential communications ports, which include Universal Serial Bus (USB) on a few types, that can be additionally utilized for loading apps out of computers. Even the micro-controllers are on average programmed with a dialect of characteristics from programming languages and C++. Besides using current compiler tool chains, the Arduino endeavor supplies a development environment (IDE) dependent around the Processing terminology undertaking [10].

From the full earth, GSM can be popular utilize digital cell anti-virus technique. GSM utilizes three leading digital radio telecommunication systems which are GSM works on 900 MHz or even 1800MHz frequency group, right after digitizing and squeezing, the GSM sends info down to a station with just two other flows of consumer data with their time slot. Together with the Development of wireless cellular telecommunications, GSM performs a critical section along with different technologies which have normal packet radio process High speed Circuit-Switched Data (HSCSD), Universal Mobile Telecommunications Assistance (UMTS) and Improved Data GSM Environment (EDGE) [11].

2 Literature Review

The internet of things (IoT) was recognized in many softwares throughout different domain names, like inside the medical industry. IoT continues to be known because of its revolution in simplifying new health together with long-term wide-range possibilities, for example, economic, technical,and societal. This analysis intends to set up IoT-based sensible home-security solutions to real life wellness tracking engineering in telemedicine structure. A multi-layer taxonomy is now driven and ran inside this review. At the very first coating, a comprehensive study on telemedicine, that centers upon the customer and host components, exhibits the other studies related to IoT-based sensible dwelling software have many limits that continue to be unaddressed. Mainly, distant patient tracking healthcare software introduces various centers and rewards by embracing IoT-based sensible dwelling technologies without undermining the stability conditions along with a potentially high quantity of dangers.

A comprehensive investigation will be executed to spot posts that cope with such topics, linked software will be reviewed, and a coherent taxonomy for those content is created. Afterward, the content articles predicated on IoT scientific tests which are connected using telemedicine software are all filtered. Six posts are chosen and categorized to just two classes. The first class that balances for 22.22% (n =2/9) comprises polls on telemedicine posts along with also their programs. The next group that accounts for 77.78% (n =7/9) comprises posts about the customer and host components of telemedicine structure. The accumulated studies show the critical necessity of building a second taxonomy coating and examine IoT-based sensible home-security research workers [12].

Even the current differences and tendencies within this region need to be researched to present invaluable fantasies for specialized surroundings as well as research workers. So, 67 content is got from the subsequent coating of the taxonomy and so are categorized into 6 classes. From the first class, 25.37% (n =17/67) of these content concentrate on the design layout. From the next group, 17.91% (n =12/67) comprises safety evaluation posts that explore the study area at the safety field of IoT-based smart house software. From the fourth group, 17.91% (n =12/67) includes safety evaluation. From the category, 13.43% (n =9/67) investigations safety protocols. From the closing group, 14.92% (n =10/67) diagnoses the safety frame [12].

Smart homes may employ new Internet of Things theories together side RFID systems for producing omnipresent products and services. This paper presents a book read out a system to get a wireless Master Slave RFID reader design of multi-standard NFC (Near Field Communication) and UHF (Ultra High Frequency) systems to construct a wiser dwelling service technique which benefits regarding expenditure, electricity intake, and sophistication. Many sensible dwelling service usage cases like washing machines, shopping, cooking, along with older wellness treatment are clarified as illustrations which use this system [13]. In this paper, an IoT-based sensible household process is constructed,which is made up of various subsystems demanded to get an intelligent residence. All these subsystems are vital parameters tracking together with attentive, security devices, power keeping approach, electric machine controller, and tracking platform. An in-depth study was accomplished to come across the proper components and applications tools to match the absolute essentials of the wise residence. Basic safety is just one among the significant factors for IoT app that continues to be addressed utilizing a protected cloud system that offers that the authentication is working with the login ID and password management technique. This newspaper suggests a comfortable, productive,and robust structure [14].

Home-automation established IoT is adaptable and hot software. In-house automation, most all dwelling appliances are all networked with each other and equipped to use with no human participation. Home-automation supplies a substantial shift in humans' life that presents smart functioning of appliances [15]. Which prompted us to build up a brand-new method which controls several appliances for the home such as lighting, supporter, door cartons, electricity ingestion, and grade of this gas tube utilizing a variety of detectors. Such as LM35, IR detectors, LDR module, and Node MCU ESP8266, along with Arduino UNO. The suggested strategy employs the detector and finds that the existence or lack of a specific thing while in the sanity consequently. Our answer also supplies advice regarding the vitality absorbed by the home proprietor from the kind of concept [16].

This outstanding issue stipulates a new medium for content articles that review mobile IoT-based CE apparatus, programs, and programs or even research novel investigation paradigms like smart houses and bright metropolitan areas. Inside this matter, discover posts on wise home surroundings, business apparatus to the border of both IoT and user cognitive programs, and newfound solitude. The prospective audience comprises teachers, investigators, and college students that are participated in IoT-based investigation and instruction [17].

3 Smart Home Master-Slave RFID System Architecture

In master-slave architecture, several readers comprised in home such an RFID reader system architecture is introduced in this paper. Smart home environment architecture is illustrated like UHF, which is one of the RFID standard protocols between tag and reader communication protocol. The system is consisting of the following reader components [18].

A Master Reader (MR)

The master reader is either direct or wireless. A stable connection could be an ordinary active static reader into a home server. Around the wise property process, require reader providers have been completed outside however in the servant reader that it starts the scanning procedure and some other inactive tags wake-up into almost any new agency initiation or even power up. Additionally, it gathers the item-level info andfor additional processing forward it into the backend. In smart homes, for advice assistance provisioning any MR can keep in touch together with almost any different MR in between remote or regional host systems along with MRFID reader may perhaps work as a proxy as shown at "Figure 1" [18].

B Several Slave Readers (SR)

Several Slave Readers are acting as relays because capturing tag ID information of the master reader by direct radio transmission is not reachable. In home appliances, Slave Readers could be integrated; when the system knew the slave reader physical location than for the localization of tags, Slave Readers location can be used.

C Master Reader (MR)

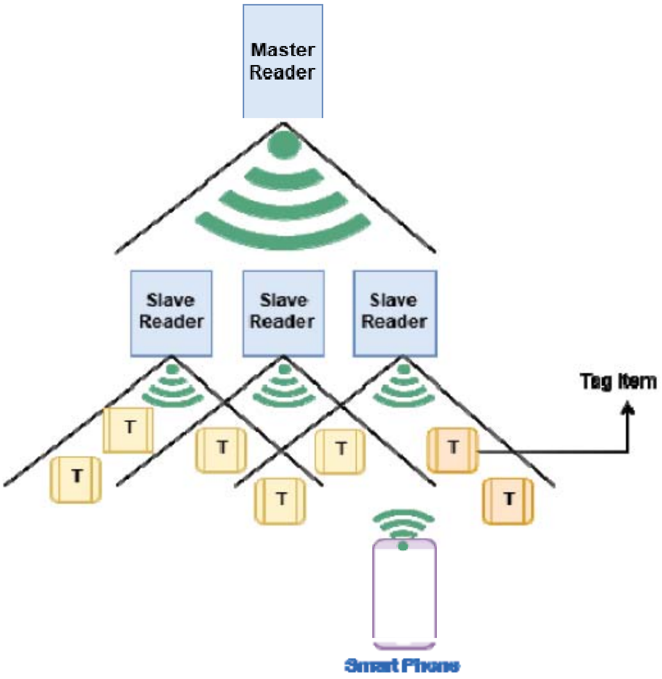


Figure 1: RFID Master-Slave Architecture Diagram

D Mobile RFID (MRFID) Reader

MRFID readers have many functionalities such as communication with remote or local servers, tag collision management, initiating the writing/reading process and waking up passive tags. These processes are very power-hungry. For 100% detection of tags, they need bigger reading window sizes depending on reading distance or human movement for MRFID reader, which implies longer active operation time. In the proposed architecture, wake up tags are energizing for operation and as "RF Energy Generator" master-slave reader act. Therefore, wakened and powered up by master or slave readers tags are always faces by MRFID reader. That is why to initiate purpose MRFID reader does not need any wake-up procedure and acts as a passive reader. For localization of tags and fault tolerance, information from tags is used. Theoretically, more responsibility could take by the slave reader in tag information processing. On the complexity and cost restrictions, slave readers intelligence is determined [18]. Also, whether mobile devices are embedding or not which capabilities of an RFID reader that can support our system from mobile devices to initiated RFID services. There can be four possibilities:

E RFID reader services for smartphones

To access RFID tag information, smart phones can be used to some extent. Communication with master-slave reader system can be done wirelessly like WLAN that is used by the mobile phone. Information about an item which is of our interest is sent through a smart phone.

F The conventional way of MRFID tag interaction and reader

In this method, as a case study user wants to know some specific item information through direct touch and with mobile RFID reader, the user equipped the information. For a small operation, this method works very well, but as mentioned earlier, with multiple, it can be mighty and time-consuming.

G MRFID reader and tag communication through master or slave reader

As a proxy master-slave system is used to communicate MRFID reader with tags. For example, to the master-slave reader, a search profile is sent by MRFID reader. Item of interest location is navigating or in the propinquity of MRFID reader, wakes up right tags is done by the master-slave reader. In the first case, the items answer (the item in proximity) and in listening mode information is consumed by MRFID reader. In this case, the mobile reader is initiated by the multiple tag collision processes; possibility to avoid power consumption is the advantage. Mobile reader delegation possibility is another benefit for the master-slave reader, as shown in "Figure 2"

H Master or slave reader as a power source

The main difference in this system acts for operation and wake-up of tags only as an RF energy source. Otherwise, between tags and MRFID reader, communication is as usual. Communication needs to be synchronized in MRFID and master-slave. So, only first MRFID reader sends a wake-up signal (step (6)). With the second wake-up call, the master-slave reader continues to tag synchronic signal of MRFID reader first wake-up (step (7)). The ID information of tags is sent to tags (step (8)). In listen mode, information is consuming by MRFID for the navigation and automatic identification tagged.

I RFID Reader Service Communication

The identification process is initiated towards the RFID reader system through smart spaces identification service application: (1) in "Figure 2". Responsibility is taken by a master reader in radio range of its wake-up tags and slave readers in step (2). In the same way, slave readers proceed to step (3) if tags are not woken up by the first wake-up call of master readers then in their range wake-up the tags. Master reader, identification information of slave tags and readers, is sent by its own, and that is a step (4), for further processing sends to smart spaces object server finally, but before it, redundant information is collected and removes in step (5). With device identification tag MRFID tag is equipped in an optimal case that is identified automatically from master-slave reader system. For this purpose, let us take an example; when it enters a shopping section identification tag with the mobile reader from section slave or master readers, it will be detected in step (9). In smart spaces, there is navigation or localization of user of the RFID reader; such additional services are allowed by this feature. From any other information servers, external objects can query by MRFID reader in all these steps (13-16).

J Smart Home NFC and RFID Services

Smart home services are based on NFC, RFID,or both.

1) NFC services

RFID, NFC, or both can be based on smart home services. In an application like (medication and presence) shopping and care of older adults, NFC services can be used. The user makes charts of their exciting items for shopping and home appliances like shelves, fridge and so on and plugs them. For checking the availability of cooking items,users visit shelves or fridge. With an appliance, a chart of items is attached to it and loaded from it in case of missing items. It is attached to a plastic cover, or A4 paper NFC tags are charts of an array. Printed picture of items with the A4 paper is fixed on top of the plastic cover.

2) RFID based services

In this case, it is much easier to control the stack of the item, because multiple tags simultaneously can be read by UHF readers. The user opens shelves or fridge, and all items can be read by just one touch. With the cooking recipe, a stack list is controlled by shopping application and automatically generates a shopping list, and this is the combined service of NFC and RFID.

3) RFID and NFC combined services

To find items potential location,this solution helps us very well, for example; in shelf NFC tag, items with their location are associated in this case. Before putting items inside, ID of the shelf is read by mobile NFC reader. While placing it on the shelf, every item is read from NFC reader and with this application items IDs and shelf are associated together. During search service, all items can read by an RFID reader at once and from the previous step using location context, read items can localize by the application as shown in “Figure 2”.

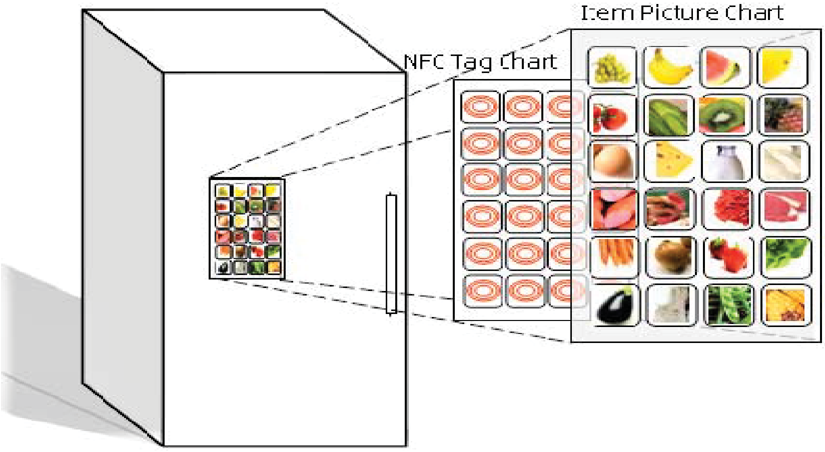


Figure 2: Fridge item chart with NFC tags[18]

4 Block Diagram for Real-Time Simulation Using RFID

In the proposed architecture diagram of the smart home system and services, we can see that our smart home devices or tag items like washing machine, air conditioner,and so on relates to the RFID reader. This reader relates to the Arduino UNO that will send the SMS that is transferred by the GSM module by using the wireless medium, this SMS or information will be sent on any Android device so users can collect the information from any tagged item in the smart home more easily and precisely as shown in the “Figure 3”.

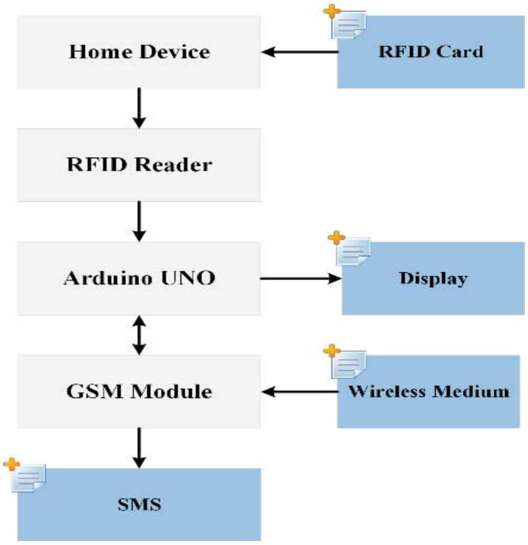


Figure 3: Architecture Diagram of smart home system and service

5 Results and Discussion

A Hardware component

1. Android Device as Virtual Terminal
2. TWN413.56MHz is an RFID/NFC reader
3. Arduino UNO
4. GSM
5. Tag Item or Home Device

B Real-time simulation using Proteus Simulator

The diagram illustrates the working of a Smart home with an RFID reader. We have used Arduino UNO with TWN4 RFID reader based on ATmega328P. Arduino UNO is a micro-controller board, it has 14 digital I/O pins, and 16 MHz clock speed whereas TWN4 RFID reader is a super small, well established and highly flexible RFID that is capable of dynamically reading a broad spectrum of radio frequencies and tags. This reader is a multi-frequency reader that nearly supports all types of transponders in one reader. It can communicate with UART (TTL, RS232), CAN, ETHERNET, USB, 12C, Clock/Data, Wiegand and SPI. We have 8 GPIO pins on TWN4 RFID reader that can directly be connected to the Arduino UNO. For Simulation, we have used the serial port where Rx is connected to the Rx of Arduino and Tx is connected to the Tx of Arduino. The circuit is shown in “Figure 4”.

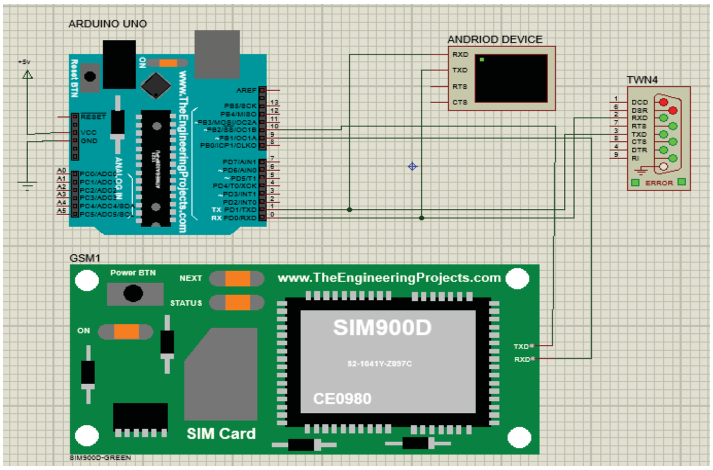
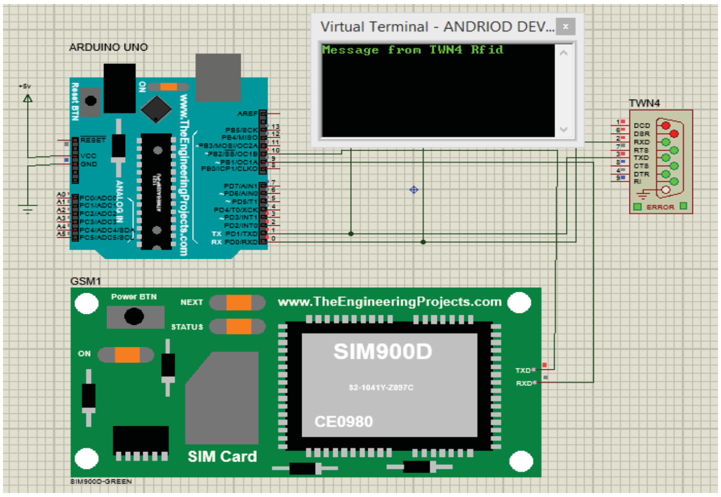


Figure 4: Simulation diagram before receiving SMS

That is a general simulation of TWN4 for washing machine, refrigerator and so on. The RFID reader is attached to Arduino, and this is attached to a specific device. When the device wants to send a message, it will use GSM to send a message to an android device. For example, RFID tags are attached to clothes that will represent the information about the material of clothes, suitable washing program and the color of the clothes. In the smart home, washing machine includes a TWN4 reader in dirty clothes container, clean clothes shelves, and the washing machine. So, this smart home application while washing clothes can monitor the number of clothes with an RFID reader and generate alarms automatically when amount of clothes reaches a specific threshold value. This machine works on energy-aware washing program identifies the time to wash clothes and sends message to the Android device. Also, the system automatically checks if some clothes in the machine are still left dirty then the machine informs users that some clothes are still dirty. This whole result is shown in "Figure 5".



```
Message from TWN4 Rfid
```

Figure 5: Simulation diagram after receiving SMS

C Use cases of IoT based smart home system and service using RFID

1) Water Tap

RFID tags can be attached to the water taps of bathroom and nay tap in the smart home; if the tap of water remains open, then this RFID reader will send SMS on the Android phones that the tap of water is open or leaked, so the user can readily do some act for this information. There could be a severe scenario that little children play freely in the home so there could be a chance that they drown in water of that child. So using IoT, we can inform users that their tap is open, by this user will be saved from a significant loss and human life could be saved.

2) Washing machine

RFID tags are attached to clothes that will represent the information about the material of clothes, suitable washing program and the color of the clothes. In the smart home washing machine includes an RFID reader in dirty clothes container, clean clothes shelves, and the washing machine. So, this smart home application in regard of washing clothes can monitor the number of clothes with an RFID reader and generates alarm automatically when amount of clothes reaches a specific threshold value. This machine works on energy-aware washing program that identifies the time to wash clothes and sends message to the Android device. Also, the system automatically checks if some clothes in the machine are still left dirty then the machine informs the users that some clothes are still dirty.

3) Kitchen

In the use case of the kitchen, we have used RFID and internet services in the cooking. It will propose a food recipe based on strong preference and other requirements like healthcare and wellness. The tags in the oven, shelves, gas cylinder and fridge communicate with RFID reader in the kitchen. In the area of shopping, list user interacts with the server for the latest item stack situation.

4) Aquarium

There should be tags on the transparent side in every aquarium where animals or aquatic plants are kept and displayed. The information we get from the tags inside the aquarium is oxygen, food availability and water replacement.

5) Furniture

In the smart home, we can attach tags with the furniture as we can get information from them. Information could be the location of these dummy things in our home. As a man working in agencies, he needs information from all dummy things because, intelligence companies are very possessive about the unique information from all the things.

6 Conclusion and Future work

IoT predicts the interconnection between applications and humans also can interconnect billions of things by using a new way of the machine to machine communications. On behalf of the user's things can interact with each other automatically, rather than always interacting with the users. This paper is based on the technology of Smart home system using the RFID reader, which is a cost-effective and energy efficient system. Using this RFID, human life can be safe and get more fruitful results. We have done an extensive simulation for this, as shown in the paper above. This system solves the serious power consumption problem of a current mobile RFID reader for technologies like RFID/UHF (Ultra High Frequency). The result is also shown in a form that RFID sends an SMS to Android device. This technology provides us with remuneration in terms of complexity, energy consumption,and costs to save human lives. Many smart home use cases such as water taps, washing machine, Kitchen and Gas cylinder so on are described as examples that make use of this system.

In the future, we can enhance the efficiency and can develop more cost-effective system by using latest Multi-standard systems and enhance the security of the system more precisely. They are also building blocks of IoT. IoT represents a confluence of many modern technologies working together namely hardware, sensors, cloud computing and machine learning, more data, better outcomes, smart, physical workspaces, safer places to work, increased customer centricity, industry revenue impact.

References

[1] L. Coetzee and J. Eksteen, "The Internet of Things-promise for the future? An introduction," in IST-Africa Conference Proceedings, 2011, 2011, pp. 1-9: IEEE.

[2] M. Elkhodr, S. Shahrestani, and H. Cheung, "A smart home application based on the Internet of Things management platform," in Data Science and Data Intensive Systems (DSDIS), 2015 IEEE International Conference on, 2015, pp. 491-496: IEEE.

[3] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A survey on wireless indoor localization from the device perspective," ACM Computing Surveys (CSUR), vol. 49, no. 2, p. 25, 2016.

[4] B. A. Ali, H. M. Abdulsalam, and A. J. W. P. C. AlGhemlas, "Trust Based Scheme for IoT Enabled Wireless Sensor Networks," vol. 99, no. 2, pp. 1061-1080, 2018.

[5] M. Elkhodr, S. Shahrestani, and H. Cheung, "A contextual-adaptive location disclosure agent for general devices in the internet of things," in Local Computer Networks Workshops (LCN Workshops), 2013 IEEE 38th Conference on, 2013, pp. 848-855: IEEE.

[6] S. Mayara Sousa and S. Raissa Boeng, "Um estudo sobre a utilização do NFC: Tecnologia que tende a aproximar a Internet das Coisas da vida dos brasileiros," 2017.

[7] H. Sugiyama and K. Nosu, "MPPM: A method for improving the band-utilization efficiency in optical PPM," Journal of Lightwave Technology, vol. 7, no. 3, pp. 465-472, 1989.

[8] L. Liu, M. Zhang, M. Liu, and X. Zhang, "Experimental demonstration of RSOA-based WDM PON with PPM-encoded downstream signals," Chinese Optics Letters, vol. 10, no. 7, pp. 070608-070608, 2012.

[9] S. Arduino, "Arduino," Arduino LLC, 2015.

[10] M. McRoberts, Beginning Arduino. Apress, 2013.

[11] M. Mouly, M.-B. Pautet, and T. Foreword By-Haug, The GSM system for mobile communications. Telecom publishing, 1992.

[12] M. Amadeo, A. Giordano, C. Mastroianni, and A. Molinaro, "On the integration of information centric networking and fog computing for smart home services," in The Internet of Things for Smart Urban Ecosystems: Springer, 2019, pp. 75-93.

[13] H. Ning, F. Shi, T. Zhu, Q. Li, and L. Chen, "A novel ontology consistent with acknowledged standards in smart homes," Computer Networks, vol. 148, pp. 101-107, 2019.

[14] M. Tao, J. Zuo, Z. Liu, A. Castiglione, and F. Palmieri, "Multi-layer cloud architectural model and ontology-based security service framework for IoT-based smart homes," Future Generation Computer Systems, vol. 78, pp. 1040-1051, 2018.

[15] M. Bansal, I. Chana, and S. Clarke, "Enablement of IoT based context-aware smart home with fog computing," in Fog Computing: Breakthroughs in Research and Practice: IGI Global, 2018, pp. 251-263.

[16] L. Yao et al., "WITS: an IoT-endowed computational framework for activity recognition in personalized smart homes," Computing, vol. 100, no. 4, pp. 369-385, 2018.

[17] H. Ghayvat, S. Mukhopadhyay, X. Gui, and N. Suryadevara, "WSN-and IOT-based smart homes and their extension to smart buildings," Sensors, vol. 15, no. 5, pp. 10350-10379, 2015.

[18] M. Darianian and M. P. Michael, "Smart home mobile RFID-based Internet-of-Things systems and services," in Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on, 2008, pp. 116-120: IEEE.

The Analysis on the usage of the Video Conferencing Rooms using Classification

Arfa Hassan[1] Salma Aftab[2] Raheela Khan[3] Hira Asim[4]

Abstract

The ways and means of collaboration have been changed over the last few decades. They have extended beyond interaction within the same meeting room. Nowadays, all multinationals are installing video conference rooms in their offices globally in order to collaborate with their clients online to save travel cost and time. These video conference rooms are meant to capture the huge amount of data. Keeping in view the growth of the data in this situation, we performed an analysis on the usage of video conferencing rooms using data mining techniques. The data have been taken from a Norway based company named Cyviz . The data set is then further preprocessed and analyzed. Data reduction and data transformation have been done on the selected attributes to get better and appropriate results. A well-known data mining tool named WEKA[5] is used to perform the classification on the dataset taken into consideration. Classification algorithms named Naïve Bayes and Random Tree are applied to the dataset after preprocessing and their results are compared and analyzed. This study is an effort to analyze the usage of the video conference room so that appropriate usage of the resources can be ensured.

Keyword: Video Conference, Audio Conference, Naive Bayes, Random Tree, Classification, Data Mining, WEKA

1 Introduction

Cyviz is a technological organization which is based on research and development. They are having 120 employees who are operating globally. It has customers in 50 countries around the world. Cyviz started back in 1999 and with its 20 years of experience they now develop and produce softwares and hardwares that are used in collaboration systems and visualization. This comprises huge display walls with high resolution and collaboration rooms that enable consumers a proficient use of display walls. Previously, the users had ability to use display walls for presentations and high resolution videos while Cyviz introduced the flexible control system which is called CDC (Cyviz Display Controller) for the concept of the collaboration rooms. It enables consumers to utilize the same display wall for sharing presentations and video conferencing with other users. In advance, there was an only one point of control for display walls.

The CDC enables to configure the admittance of numerous consumers to control the system. The 20 years of focus and dedication on the technology of large displays has given

[1] *University of Management and Technology, Lahore | f2017114004@umt.edu.pk*
[2] *University of Management and Technology, Lahore | s2018114003@umt.edu.pk*
[3] *University of Management and Technology, Lahore | s2018179006@umt.edu.pk*
[4] *University of Management and Technology, Lahore | hira.asim@umt.edu.pk*

[5] *WEKA Website: https://www.cs.waikato.ac.nz/ml/weka/*

Cyviz a unique perception which enables them to understand the requirements of different businesses. In order to improve the portfolio of their product the company is determinedly spending in new technology [1].

Cyviz has provided meeting rooms to the Fortune 500 and the government customers. These meeting rooms have different features and they are installed according to the customer's requirement. These features are providing information about the usability of the conference rooms. The features of video conferencing rooms are like audio detected, room in use, the room booked through exchange, presence detected, picture in picture active, stereo on, audio or video conference active, audio call active, video call active and some others.

This study is based on the analysis of available features of the conference rooms and extracting meaningful information out of it.

2 Background Knowledge

Video conferencing is the means of communication using a combination of audio, video, text and graphics to support real time communication between distributed groups sharing same interests or working in the same domain like business meetings, playing games, learning and entertainment [4]. Our study is a statistical analysis of a system using video conferencing determining the availability and non-availability of rooms using the features provided by the organization. Weka is an application that runs almost on every system and is developed using JAVA. It provides an interface to multiple algorithm and also support pre and post processing of data to extract results from different data sets [4]. Random trees are used to predict results using multiple decision trees that grow in different subsets in the same domain, the idea was proposed by Leo Breiman in 2000 [5]. Naïve Bayes is a simple classifier that calculates the probabilities of frequency and uses different combination of values from a data set, the algorithm is helpful in supervised learning [6]. Depending upon our best knowledge and research study, we tried to analyze the data set received from Cyviz using above mentioned data mining techniques in order to find meaningful knowledge.

3 Methodology and Results

The machine learning view of Knowledge Data Discovery (KDD), shown in the Figure 1, is followed in the whole research process.

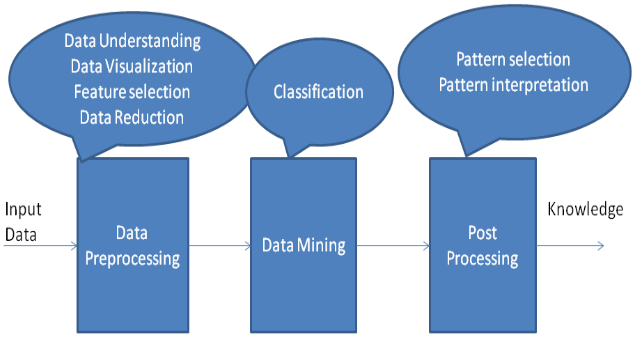


Figure 1: KDD process a machine learning view

A Data Preprocessing

The phase is handling the data manipulation Procedure of data preprocessing consists of data understanding, data visualization, feature selection and data reduction.

1) Data Understanding

The research started from getting familiarity and understanding of the available data. The initial data provided for analysis were having around 30 features of the video conference rooms. These features were capturing the usage time in number of seconds, more precisely the number of seconds each feature had been used in any particular room. For example, in one transaction for any particular date if Audio conference feature is having 3600 value, then the room used the feature of audio conference for an hour. The features were studied thoroughly to figure out their usage and any possible relation.

2) Data Visualization

In order to get more understanding of the data all available features were plotted in MATLAB. Some of the plots that are understandable are shown in Figure 2, Figure 3 and Figure 4.

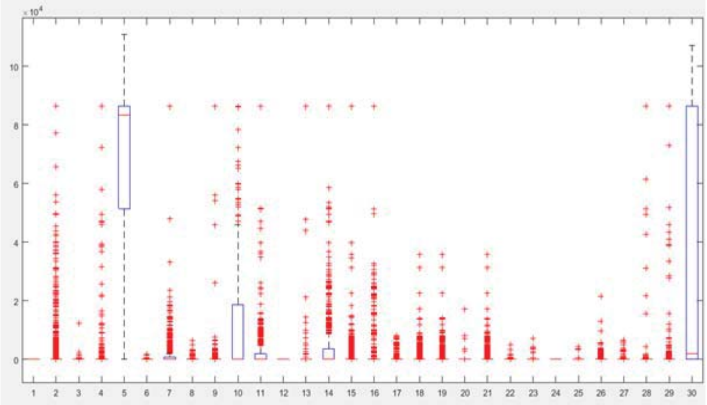


Figure 2: Boxplot showing all features of the video conference room

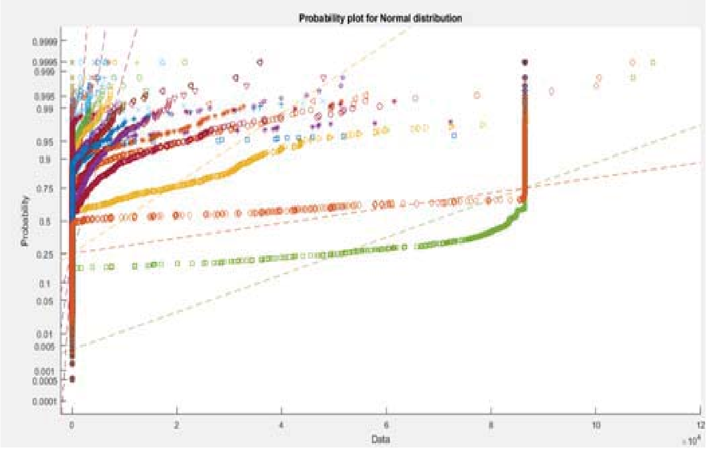


Figure 3: Probability plot for Normal distribution

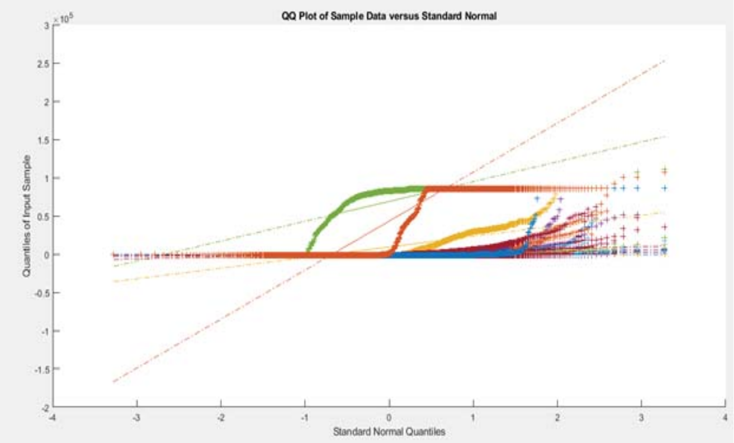


Figure 4: QQ Plot of sample data versus standard Normal

3) Data Reduction

From the visualization of data, it was figured out that in order to understand the patterns which were more informative, it was important to reduce the data size by selecting some maximum information giving features. The selected features of the study are shown in Table 1.

Table 1: The selected feature of video conference room

Feature	Description
Room Name	The name of the room. The name uses name space notation, showing the region, country and city of that room.
Day	The day of the week on which this information is recorded. The data is captured against particular date and day is derived from that date.
Is Room In Use	The number of seconds for which room was in use on a particular day
Is Pip Active	The number of seconds for which picture in picture was active on a particular day
Is Audio Video Conference Active	The number of seconds for which audio and video conference both were active. It has sum of Audio and video active time collectively on a particular day

4) Statistical Description

The statistical description of selected features is shown in the Table 2 below.

Table 2: The usage data in number of seconds

Description	Is Room In Use	Is Pip Active	Is Audio Video Conference Active
Mean	5114,80129	2627,081	1500,87895
Median	0	0	0
Mode	0	0	0
Standard Error	349,335813	229,8189	201,5143526

Standard Deviation	15094,3797	9930,199	8707,192431
Minimum	0	0	0
Maximum	89999	86400	86400

The average use of all rooms together is 85 minutes.

Mean: = 5114,80129/60 = 85 minutes

The average use of Picture in picture for all rooms is 44 minutes.

Mean: = 2627,081/60 = 44 minutes

The average use of audio and video conference feature for all rooms is 25 minutes

Mean: 1500,87895/60 = 25 minutes

5) Data Transformation

In order to apply data mining algorithms on all the selected features, they were transformed. The transformed data are shown in Table 3.

Table 3: The data before and after transformation

Feature	Actual Data	Transformed Data
Room Name	Number of seconds	Nominal
Day	Day against each date	Day
Room Used	Number of seconds	Boolean
Picture in picture	Number of seconds	Boolean
Audio or video conference	Number of seconds	Boolean

The main objective of data transformation was to generate meaningful knowledge after applying data mining techniques. The rules that were followed for transformation are defined in Table 4.

Table 4: The rules set for data transformation

Feature	Actual Data	Transformed Data
Room Used	= 0 > 0 &<= 120 mins >120 &<= 360 mins >360 mins	No Usage Low Usage Normal High
Picture in Picture	= 0 > 0	N Y
Audio or Video conference	= 0 > 0	N Y

Table 4 above is giving us the following insights about the features we took into consideration:

a) Room used

If there is no meeting at all then recorded data is 0 seconds, which is transformed into No usage of room. If the room is used for 1 minute to 2 hours (120 mins) then the usage of the room is Low. If the room is used for 2 hours to 6 hours (360 mins) then the usage is Normal. For more than 6 hours the usage is transformed into High.

b) Picture in Picture sharing

If the data of picture in picture sharing feature are 0 then it is transformed into N otherwise it is Y.

c) Audio and video conference

If the data of audio and video conference feature is 0 then it is transformed into N otherwise it is Y.

B Data Mining

1) Data Sample

There are 6 attributes and Room used is the class attribute. The total number of transactions is 2449. The sample data is shown in Table 5.

Table 5: The sample dataset after preprocessing

Room	Day	Room Used	PIP	Audio Video	Usability
Room 1	Tuesday	No Usage	N	N	no usability
Room 1	Wednesday	Low	Y	Y	full utilized
Room 1	Thursday	Low	Y	N	min-use
Room 1	Monday	Normal	Y	Y	full utilized

2) Weka Tool for Classification

Weka is providing different machine learning algorithms for the tasks of data mining [4]. The research classifiers are implemented in it.

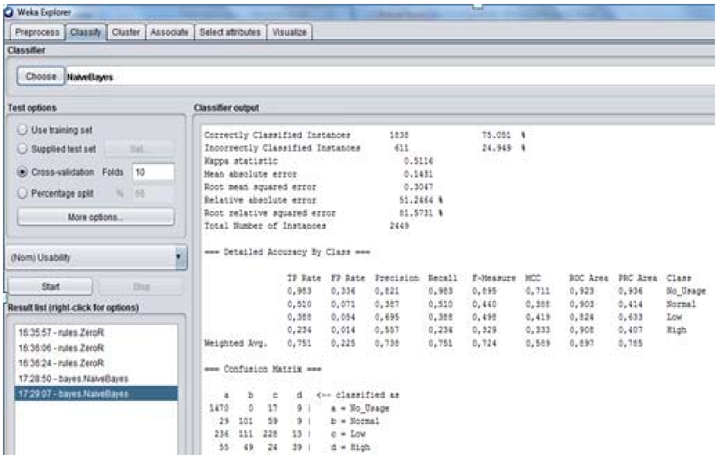
3) Naïve Bayes Classification

The Naviebayes classifier is based on Baye's theorem of posterior probability. It is assumed that features or attributes are independent of each other. The value of one attribute is independent of the other attribute of a class [2]. The summary of accuracy generated by Naive Bayes Classifier is shown in Table 6.

Table 6: The summary of accuracy by Naive Bayes

Correctly Classified Instances	1838	75.051%
Incorrectly Classified Instances	611	24.949%

The implementation of Naive Bayes Classifier is shown in the Figure 5.



Classifier: NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Usability

16:35:57 - rules.ZeroR
16:36:06 - rules.ZeroR
16:36:24 - rules.ZeroR
17:28:50 - bayes.NaiveBayes
17:29:07 - bayes.NaiveBayes

```
Correctly Classified Instances        1838               75.051  %
Incorrectly Classified Instances       611               24.949  %
Kappa statistic                          0.5116
Mean absolute error                      0.1431
Root mean squared error                  0.3047
Relative absolute error                 51.2464 %
Root relative squared error             81.5731 %
Total Number of Instances             2449

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,983    0,336    0,821      0,983    0,895      0,711    0,923     0,936     No_Usage
                 0,510    0,071    0,387      0,510    0,440      0,388    0,903     0,414     Normal
                 0,388    0,054    0,695      0,388    0,498      0,419    0,824     0,633     Low
                 0,234    0,014    0,557      0,234    0,329      0,333    0,908     0,407     High
Weighted Avg.    0,751    0,225    0,738      0,751    0,724      0,569    0,897     0,785

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1470    0   17    9 |    a = No_Usage
   29  101   59    9 |    b = Normal
  236  111  228   13 |    c = Low
   55   49   24   39 |    d = High
```

Figure 5: Implementation of Naive Bayes using WEKA

The confusion matrix using Naive Bayes classifier is shown in the Table 7.

Table 7: The confusion matrix using Naive Bayes classifier

A	B	C	D	Classified As
1470	0	17	9	A = No Usage
29	101	59	9	B = Normal
236	111	228	13	C = Low
55	49	24	39	D = High

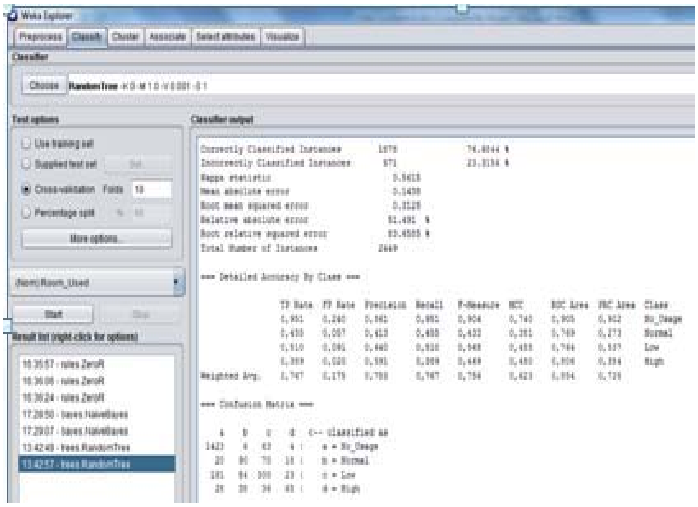
4) Random Tree Classification

Random tree is an ensemble machine learning algorithm which generates numerous different learners. In order to produce a random set of data for constructing a decision tree it uses a bagging idea [4]. An alternative classifier is used in the research study to get better accuracy. The summary of accuracy generated by the random tree is shown in the Table 8.

Table 8: The summary of accuracy by random tree

Correctly Classified Instances	1878	76.684%
Incorrectly Classified Instances	571	23.316%

The implementation of Random tree classifier is shown in the Figure 6.



RandomTree -K 0 -M 1.0 -V 0.001 -S 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Room_Used

10:35:57 - rules.ZeroR
10:36:06 - rules.ZeroR
10:36:24 - rules.ZeroR
17:28:50 - bayes.NaiveBayes
17:29:07 - bayes.NaiveBayes
13:42:49 - trees.RandomTree
13:42:57 - trees.RandomTree

```
Correctly Classified Instances        1878               76.6844 %
Incorrectly Classified Instances       571               23.3156 %
Kappa statistic                          0.5613
Mean absolute error                      0.1435
Root mean squared error                  0.3125
Relative absolute error                 51.491  %
Root relative squared error             83.4565 %
Total Number of Instances             2449

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,951    0,240    0,861      0,951    0,904      0,740    0,905     0,902     No_Usage
                 0,455    0,057    0,413      0,455    0,433      0,381    0,769     0,273     Normal
                 0,510    0,091    0,640      0,510    0,568      0,488    0,764     0,537     Low
                 0,389    0,020    0,591      0,389    0,469      0,480    0,806     0,394     High
Weighted Avg.    0,767    0,175    0,758      0,767    0,756      0,623    0,854     0,729

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1423    6   63    4 |    a = No_Usage
   20   90   70   18 |    b = Normal
  181   84  300   23 |    c = Low
   28   38   36   65 |    d = High
```

Figure 6: Implementation of Random Tree classifier using WEKA

The confusion matrix generated in WEKA using the random tree classifier is given below in Table 9.

Table 9: The confusion matrix using Random Tree classifier

A	B	C	D	Classified As
1423	6	63	4	A = No Usage
20	90	70	18	B = Normal
181	84	300	23	C = Low
28	38	36	65	D = High

C Post Processing

The proposed classifiers have divided the data into four categories No usage, low, medium and high room usage. They are giving two different accuracies and can predict the usage of the room on the basis of defined attributes.

4 Conclusion & Future Research Directions

The selected features of the video conference room are showing the level of utilization of that particular room. In future, work we are interested to figure out that on which day, week or a month meeting room is busy or it can predict about the availability of the room on any particular day or week. As the company is ready to provide more data for analysis so we have a room to expand our work in different directions. We also suggest the company to keep the record of the camera installed in the room in order to check the correct usage of a conference room. Through camera we can obtain image data and can relate this to the existing features.

Acknowledgment

We acknowledge the support of Cyviz for providing us their valuable data for analysis and the continuous guidance of Dr. Malik Tahir Hassan.

References

[1] Cyviz. (n.d.). Retrieved Febraury 01, 2019 from Cyviz: https://www.cyviz.com/about/

[2] Jiawei Han, M. K. (2012). Data Mining Concepts and Techniques. Waltham: Morgan Kaufmann.

[3] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

[4] Kalmegh, S. R. (2015). Comparative analysis of weka data mining algorithm randomforest, randomtree and ladtree for classification of indigenous news data. International Journal of Emerging Technology and Advanced Engineering, 5(1), 507-517.

[5] Benedictus, F. (2011). Get more out of video conferencing (Master's thesis, University of Twente).

[6] Patil, T. R., &Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

Lean Product Development (LPD)-A Systematic Literature Review

Rida Fatima[1] Khubaib Amjad[2]

Abstract

Context: The major goal of development processes in software engineering is to avoid unnecessary features and to provide value driven software products. Lean product development (LPD) resolves these issues to some extent by emphasizing on reducing waste and increasing customer value. LPD is reaction based approach. The ultimate goal of lean is to eliminate waste, balance the process and to undergo demand-driven planning. Lean brings flexibility in the development and manufacturing processes. Lean works on its seven principles hence improving the development process. Objective: The main objective of this systematic literature review (SLR) is to get insight of lean product development. We aim at identifying, summarizing, and analyzing the existing high-quality primary studies on principles of Lean and discuss its positive and negative impacts. Lean's impact on various aspects of software development organizations (SDO) such as people, process and product is analyzed. The methods used by different organizations to apply lean principles to software development processes along with their models, tools and frameworks are discussed. There are several studies on LPD and there is no systematic review performed so there is need of an effective and unbiased SLR in this domain. Methodology: The selection process includes data extraction, detailed analysis and reporting of findings. Primary studies are selected by following a systematic and unbiased selection procedure according to standard PRISMA guidelines. This study highlighted five research questions which need to be addressed regarding LPD. Conclusion: Lean principles have their impact on all stages of development employing its seven principles. Chronological distribution of selected studies have shown a decreasing trend of LPD after 2014 which also justifies the need for this systematic literature review. Findings include the methods used for LPD so far and overall research productivity of this domain. This SLR also discusses more research perspectives to be considered in this domain.

CCS Concepts: • Software and its engineering • Agile software development.

Keywords: Lean Product Development, Lean Principles, Lean Management.

1 Introduction

Lean product development (LPD) has its roots in the Japanese automotive industry from many years but now it has produced a significant impact in manufacturing environments to improve organization's performance. Due to its main focus on reducing waste and increasing customer value, lean has also found its way in other domains such as healthcare, government and service industries[1]. Lean is not only confined to product development and technical aspects rather it also has its soft practices i.e. impact on human factors, business and managerial activities of an organization which are even more influential towards lean success[2]. Lean concepts have their

[1] *National University of Computer & Emerging Sciences, Islamabad | f179023@nu.edu.pk*
[2] *National University of Computer & Emerging Sciences, Islamabad | khubaib.amjad@nu.edu.pk*

origin in the Toyota production system[3]. This system regarded lean thinking as the core for developing subsystems of tools, technologies and processes [4]. LPD was introduced in Toyota Motor corporation in 1950's [5]. Due to this reason, Lean concepts have been majorly applied in automobile industry so far to produce high quality, low cost and shorter time to market products. There are seven principles of Lean which different companies apply according to their own production systems. According to [3], the seven principles are defined. Principle one is to Eliminate Waste; it means to avoid unnecessary features which do not add value to the final product. Principle two is; Integrating Quality; it states early detection of defects to improve overall quality and productivity. Principle three is Creating Knowledge; it emphasizes that knowledge should be stored in a way that makes it easies for new team members to understand the project without going into initial process of learning. Principle four is Postpone Commitments; it states that irreversible decisions should be scheduled to the last possible moment when the team will have more knowledge on the subject. Principle five is Delivering Fast; it argues that product will be delivered as soon as possible if continuous feedback is taken from customer in order to avoid requirements change. This can be done by dividing project in iterations. Principle six is Respect People; it focuses on enabling the team rather than controlling them by trusting their way to work so that processes can be improved and final decisions should take into account everyone's suggestions. Principle seven is to Optimize the Whole; this principle focuses on improvement of local processes to get global advantage.

The new focus of lean is on economic, environmental and social sustainability[5]. Economic sustainability includes increased economic value due to reduction of waste. Environmental sustainability is about resource usage and social sustainability focuses on employees' needs like training and education, giving them equal opportunity, autonomy and motivation to work.

Lean manufacturing (LM) has also some negative effects along with positive ones. The negative effects and their causes are stated in [6]. It is stated that 40% of all of the projects showed negative effects. Many of these negative effects emerged due to poor management and control of project.

There is significant research carried out in LPD but the chronological distribution shows that quality research contribution in LPD has been decreased since 2014 (see Figure 1). Moreover, there is no Systematic Literature Review (SLR) to present the emerging trends in this domain. This SLR attempts to fill this gap and it will also identify more research perspectives to be considered in this domain. The timespan considered for this research is 2013-2018.

Our contribution is to identify the overall research productivity in this domain. Due to limitation of number of pages, only the research productivity is provided and some methods of Lean Product Development (LPD) are enlisted. A brief discussion is also provided highlighting some concepts related to Lean. Through a quality assessment process, this study ensures that only high-quality studies meeting certain quality scores are considered for the inclusion.

This study is organized as follows: Section 2 elaborates the detailed research methodology including all research questions, data sources and Inclusion, Exclusion & Quality criteria. Subsequently, section 3 reports results and discussion of selective research questions. Discussion is concluded in section 4.

2 Research Methodology

A Systematic Literature Review (SLR) aims at identifying, evaluating and interpreting available research related to a specific field of interest[7]. An SLR needs to follow an unbiased search plan.

We aim to fill the gap of SLR in LPD domain using SLR guidelines by Kitchenham [7]. This review process comprises of three main phases; Planning, Conducting and Reporting. These phases need to be conducted in a systematic and disciplined way (see Table 1).

Table 1: Systematic Literature Review Process

Phases	Steps
Planning	Research Objective
	Selection of Online-Digital Libraries
	Formulation of the Query String
	Definition of Inclusion and Exclusion criteria
Conducting	Study Selection
	Data Synthesis
Reporting	Proposed Plan/ Result
	Report formatting

A Research Questions

The primary research question of this SLR is: "What is overall research productivity of lean product development?" Formulated research questions are listed (see Table 2).

Table 2: Research Questions

RQ#	Research Questions
RQ1	Which methods have been used so far to apply lean principles to product development?
RQ2	What is effect of lean principles on agile development?
RQ3	Is there any reported problems while applying lean principles to product development?
RQ4	How LPD principles are improving current development paradigms?
RQ5	What is overall research productivity in this domain?

B Search Criteria

The timespan of 2013-2018 is considered to conduct this SLR. A step by step filtration process is used to extract the related studies from the databases. Firstly, the studies are acquired from the databases using manual and automated search. Secondly, studies are filtered on the basis of title and abstract considering certain defined and related keywords. Finally, full-text reading of selected research papers have been performed to further clarify the relevance of studies.

C Data Sources

Data has been gathered using both automated and manual search. Automated queries are applied on popular search engines; IEEE Xplore, ACM Digital Library, Springer and Science Direct. In springer, results have been gathered for both computer science and business & management disciplines. For manual search, Google Scholar is considered.

D Formulation of Search String

After collecting Meta search terms, following are the search strings used for respective data bases considering timespan (2013-2018).

Springer: with at least one of the words ***product* design* software "life cycle" principle develop* manage**** where the title contains ***lean***.

Science Direct: tak(lean)AND tak (product* OR design* OR software OR "life cycle" OR principle OR develop* OR manage*)

IEEE: "Document Title":lean AND ("Document Title": product* OR "Document Title":design* OR "Document Title":software OR "Document Title":"life cycle" OR "Document Title":principle OR "Document Title":develop* OR "Document Title":manage*)

ACM: where **TITLE** matches all ***lean*** and where **TITLE** matches any ***product* OR design* OR software OR "life cycle" OR principle OR develop* OR manage****

E Inclusion, Exclusion and Quality Assessment criteria

Inclusion and exclusion criteria is used to select potentially relevant studies from data sources to answer the research questions in this SLR. This criteria is applied to each selected study retrieved in the initial phase of the study selection procedure. The inclusion, exclusion (IE) and quality assessment (QA) criteria employed in this SLR is listed (see Table 3 and Table 4).

Table 3: Inclusion and Exclusion Criteria

Inclusion Criteria	
IC1	Peer reviewed articles
IC2	Articles showing effect of lean principles on development paradigms
IC3	Articles discussing lean principles
IC4	Inclusion of latest study in case of multiple studies on the same theme
IC5	Articles published during timespan (2013-2018)
IC6	Articles answering one or more research questions
Exclusion Criteria	
E1	Studies other than English language
E2	Studies having sole focus on agile development
E3	Short papers, surveys, review papers and Posters

Table 4: Quality Assessment Criteria

QC#	Quality Considerations
QC1	Is the study discussing LPD principles?
QC2	Is the study discussing method used by LPD?
QC3	Is the study having clearly specified goals and objectives related to LPD?
QC4	Is there any comparison of LPD with other development paradigms?
QC5	Is the study highlights any limitation of LPD?

F Selection of Studies

Data extraction forms are designed. Studies are selected based on their fulfillment of quality criteria and research questions answered in each respective study. The study selection procedure adopted for this SLR consists of four steps and is according to the standard PRISMA guidelines for systematic review (see Figure 1). The steps include

1) Identification

Studies are selected using manual and automated search. Queries are used for automated search on different search engines.

2) Screening

Selected studies are filtered first on the basis of relevant title and then abstract basis screening is performed to further select only relevant studies.

3) Eligibility

Screened studies are accessed for full-text to check their credibility.

4) Included

Final filtered studies are included.

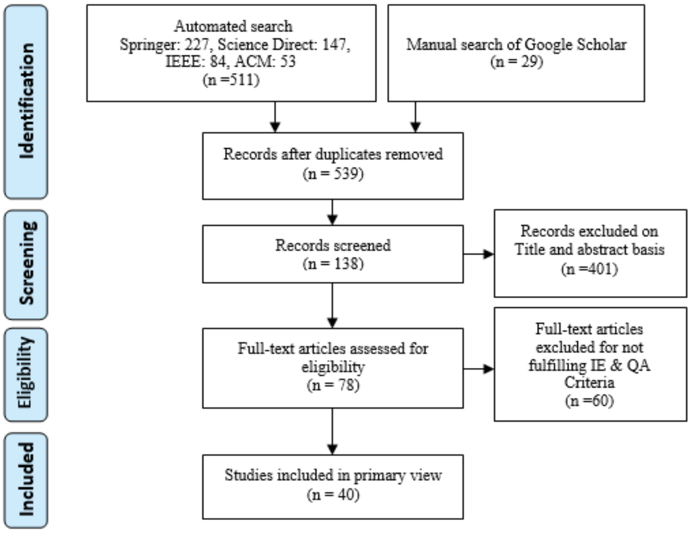


Figure 1: Study Selection Process of the SLR

3 Results and Discussion

Software development and manufacturing are two fundamentally different domains. The implementation of lean software development is explained in [8] by conducting case study in Ericsson R&D Finland and analyzing its factors of success. Also, the three challenges of lean software development are identified as achieving flow i.e. all departments of an organization should work in one direction of lean thinking, transparency and creating a learning culture which requires time and commitment of team members. This case study emphasizes that value and quality are more important concepts as compared to reduction of waste; which consequently leads to more customers and high quality.

A Lean Enabled Accelerated Planning (LEAP)

Lean Enabled Accelerated Planning (LEAP) is being discussed in [9] by conducting a case study on an international company, Rockwell Collins. This explains that planning can be done more efficiently by involving key stakeholders in up-front planning to identify risks and schedule activities to improve customer satisfaction and performance.

B Lean ERP

There is an idea of combining lean thinking with information technology in [10] to develop an Enterprise Resource Planning (ERP) based lean implementation. This study relates lean principles with ERP framework to get the desired best working system. It suggests that continuous lean learning by team is very necessary to make this new implementation successful. However, in contradiction to this, it is mentioned in [11] that lean management and ERP systems contradict in many aspects. Lean focuses on low cost, simplicity and transparent information whereas it is stated that ERP tends to be complex, costly and is not transparent. Moreover, lean is flexible to accommodate changes while ERP is inflexible because of high costs of change. So most organizations combine these two concepts which is called Lean ERP. A case study was conducted in marine sector as mentioned in [12]. It describes that the implementation of only lean tools will not suffice for achieving leanness in development process. Rather an organization must learn continuously about lean thinking to get better results.

C Impact of Lean on phases of product development

According to [13], lean principles have great impact on product development processes and tend to change the process in a drastic way. It shows that principles of lean; “Respect people” and “Optimize the Whole” have major influence on stages of development. Moreover, principle of “Integrating Quality” has an impact on product design stage.

D Waste reduction

According to [14], one kind of waste reduction is to minimize the idle time of an artifact in development process after its complete implementation. It will increase efficiency because early feedback will be available. There is a ready buffer containing most important features from which developer chooses at the moment. This scheme uses value stream mapping (VSM). The categories of wastes are mentioned in [15] which are waiting, over-production, rework, motion, over processing, inventory and transport.

E Lean Manufacturing (LM)

According to [16], there are seven dimensions of lean manufacturing (LM) stated as workforce, manufacturing process & equipment, supplier, manufacturing planning & scheduling, customer, visual information system and product development & technology. As it is mentioned in [6] that there is not even one project which is without negative effects. The negative effects imposed by LM are: 1. Late or cancelled deliveries known as Fall-outs due to internal problems, 2.Quality issues 3.Increased stock 4.Customer dissatisfaction leading to damaged reputation of company 5.Reduced sales 6. Fluctuation of core employees 7. Increased cost. The main causes of these mentioned effects are identified in the same paper which are; 80% of focus on direct effects i.e. man power and cost, 73% goes to inconsistency in planning, 70% are due to focus on whole project instead of individual iterations, 63% are related to scope i.e. not understanding it well and 63% are due to inability of determining risk at initial levels.

F Sustainability in Lean

According to analysis in [17], lean is 20% process and 80% mindset which means that in order to transform your processes to lean, every member from higher to lower hierarchy in team should be involved in process of continuous improvement. Sometimes the lean philosophy does not seem to be sustainable or undergoes failure. For this issue to handle, an assessment tool was introduced in [18] to allow companies to access lean and implement it in an effective way. There are also two more solutions proposed in [19] which are organizational memory building and institutionalizing. Organizational memory can be in the form of declarative memory including facts & events, procedural memory including procedures & functions and emotional memory of past events. This can be preserved in the form of hard data or through experts. Institutionalization means that company's principles and strategies should be stored in a way that new people will be able to learn these easily thus maintaining sustainability.

Lean principles can be categorized in terms of three dimensions; people, process and product. There are lessons learnt while transforming to lean approach considering each dimension individually provided in [28] (see Table 5).

Table 5: Lessons Learn During Lean Transformation

Dimension	Lessons Learn
Process	Guide the team but give them freedom to choose their process.
	Ensure that applied processes are right by giving the team freedom.
	Break the monotony and renew the processes.
	Focus on continuous improvement by reducing waste and adding value.
People	Keep the team members self-responsible by giving them opportunity to act as a tact speaker for a day.
	Team members should be able to pick up the knowledge on their own and there should be role rotation.
	Appreciate the whole team instead of an individual.
Product	Improve quality of internal artifacts which affect external behavior to get early investments.

RQ1: Which methods have been used so far to apply lean principles to product development?

There are many tools and methods used by different organizations to apply lean concepts in their development systems. Lean has been used in many contexts in different organizations. Lean thinking has been applied in IT service innovation, designing websites and in management perspectives.

Moreover, different systems like Manufacturing Execution System (MES) and Enterprise Resource Planning (ERP) use lean in parallel to their own functions to achieve high performance. There are certain tools proposed to access leanness of organizations and to implement lean principles in a sustainable way.

Various tools, models, methods and frameworks corresponding to different lean contexts are listed (see Table VI). Only those studies are listed in the table which have clearly defined frameworks, tools, methods and models regarding different concepts of lean. Brief descriptions are also provided for them. Their findings are listed to get insight of various lean concepts and to identify certain research gaps for future considerations.

Table 6: Analysis of Different LPD Concepts Regarding Models/Tools/Frameworks

Ref.	Concept	Contribution	Description	Finding
[1]	Use of lean principles in IT service innovation	Conceptual Framework	Analyzed the lean principles using case study of service organization	There should be enough openness between two organizations to facilitate innovation
[20]	Design of reward-based crowdfunding website	Lean Product Process Framework	Persona, Kano Model, Product Value Proposition, User Stories, User Experience Design Framework, & Usability Testing	Minimum viable product (MVP) is obtained by involving users
[21]	Lean software product management (SPM)	Erlang-C Model accompanying case study	Evidence-based decision making approach; use of Kanban in Software development	Provides decision making process for SPM
[22]	Effect of each lean principle on performance	Core benefits of each lean principle	Lean enablers are mentioned corresponding to each principle & Enablers Enablers are mapped to lean metrics (implementation & Program) to measure performance	Give suggestions of what to measure to check performance
[16]	Lean Management (LM) Dimensions	Conceptual model	Model has input, transformation & output phase	Analyzed that there are 7 dimensions of LM

[23]	Achieve effective leanness in development process	Compact teams (CTs) model accompanying case study	CTs differs from traditional PD in terms of team size, functional organization & No. of projects assigned to a designer	Significant performance benefits
[24]	Manufacturing execution systems (MES) support for lean production	5 stage capability maturity model (CMM)	Describes that how MES can be used to support lean production principles	Shows that CMM has its influence on both practical implementation & theoretical knowledge
[25]	Lean service management	Lean management Framework	Framework has five phases. Each phase having 3 principles	Analysis helps service companies to apply lean management in their operational business
[26]	Effect of human factor in lean management	Research model	Model takes individual characteristic as its input	To achieve long-term performance of lean, individuals should be given attention regarding both technical and soft practices
[6]	Assessing negative side effects of lean management	Multi-perspective assessment method	Identify negative effects and their root causes	This method can monitor, detect and overcome negative side effects
[18]	Implement lean in a sustainable way	Lean assessment tool	It has 24 criteria having 4 important factors; culture, leadership, knowledge and process. These are further sub-divided.	Certain countermeasures help to assess leanness but they vary from company to company
[27]	Assessing company's structure before applying lean	Lean product & process development performance measurement tool	Provides companies with a readiness framework to access their status before transforming to lean	Provides a sufficient framework to access development practices

RQ2: What is overall research productivity in this domain?

This is the main research question focusing on determining overall research productivity in Lean Product Development (LPD) domain so that future research will be made easier because the influential studies of this domain have already been highlighted. To answer this research question, chronological distribution of selected primary studies having time span (2013 to Jan-2018) is plotted (see Figure 2). This distribution shows that most of the studies on LPD are being in year 2013 and 2014, whereas same trend goes for both 2015 and 2016 having six significant studies each. There are four primary studies of LPD in 2017 and only one in 2018. So, 2013 and 2014 can be regarded as most productive years following 2015 and 2016. It can be clearly seen that the research interest in LPD has been decreased over the years which also justifies the need for systematic literature review in LPD. The result are shown after performing qualitative analysis in which we have applied the inclusion, exclusion and quality criteria mentioned in research method.

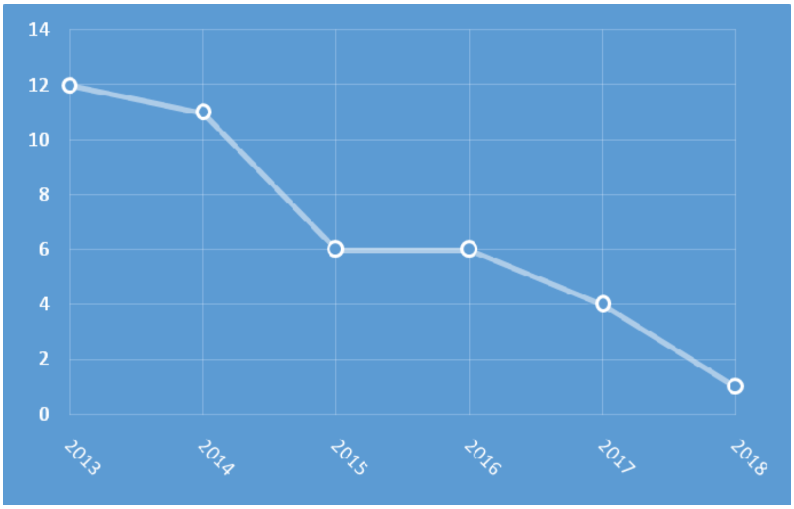


Figure 2: Chronological Distribution (2013-2018)

4 Conclusion

Lean product development is very effective to improve overall quality of development. Lean principles have their impact on all stages of development employing its seven principles. Lean surpasses other development processes by avoiding unnecessary features and only focusing on those as demanded by the customers thus increasing customer value. It delivers as fast as possible by working in iterations to get immediate feedback so that changes can be avoided thus reducing overall cost. Overall research productivity of this domain has been analyzed. In this SLR, lean product development has been discussed considering certain contexts. Moreover, methods used for lean development by different organizations in various contexts have been listed. In the extended version of this SLR, details will be broadened and remaining research questions will be given comprehensive consideration.

Acknowledgment

The authors are grateful for the funding support from the National University of Computer & Emerging Sciences (FAST-NUCES), Islamabad, Pakistan.

References

[1] Y. Gong and M. Janssen, “The Use of Lean Principles in IT Service Innovation: Insights from an Explorative Case Study,” 2014, pp. 58–69.

[2] M. F. Van Assen, “The moderating effect of management behavior for Lean and process improvement,” 2018.

[3] M. Misaghi and I. Bosnic, “Lean Mindset in Software Engineering: A Case Study in a Software House in Brazilian State of Santa Catarina,” 2014, pp. 697–707.

[4] Y. Wang et al., “Application Research of Lean Thinking in the Birth Process of Product,” in Proceedings of 20th International Conference on Industrial Engineering and Engineering Management, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 313–321.

[5] B. Resta, S. Dotti, P. Gaiardelli, and A. Boffelli, “Lean Manufacturing and Sustainability: An Integrated View,” 2016, pp. 659–666.

[6] A. M.-I. I. C. on A. in and undefined 2017, “A Method of Multi-perspective Assessment of Lean Management,” Springer.

[7] S. E. Group, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” 2007.

[8] P. Rodríguez, K. Mikkonen, P. Kuvaja, M. Oivo, and J. Garbajosa, “Building lean thinking in a telecom software development organization: strengths and challenges,” Proc. 2013 Int. Conf. Softw. Syst. Process - ICSSP 2013, p. 98, 2013.

[9] D. Secor, S. Lucae, and E. Rebentisch, “Planning for resilient lean programs,” Procedia Comput. Sci., vol. 28, no. Cser, pp. 138–145, 2014.

[10] D. Powell, E. Alfnes, J. O. Strandhagen, and H. Dreyer, “The concurrent application of lean production and ERP: Towards an ERP-based lean implementation process,” Comput. Ind., vol. 64, no. 3, pp. 324–335, 2013.

[11] M. Adam, J. Keckeis, P. Kostenzer, and H. Klepzig, “Lean erp: How erp systems and lean management fit together,” Lect. Notes Inf. Syst. Organ., vol. 4, pp. 13–18, 2013.

[12] E. L. Synnes and T. Welo, “Applicability of lean product development to a company in the marine sector,” IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2017–Decem, pp. 2281–2285, 2018.

[13] U. Dombrowski, K. Schmidtchen, and P. Krenkel, “Impact of lean development system implementation on the product development process,” IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2015–Janua, pp. 1462–1466, 2014.

[14] T. Lehtonen, T. Kilamo, S. Suonsyrja, and T. Mikkonen, “Continuous, Lean, and Wasteless: Minimizing Lead Time from Development Done to Production Use,” Proc. - 42nd Euromicro Conf. Softw. Eng. Adv. Appl. SEAA 2016, pp. 73–77, 2016.

[15] P. J. A. Reusch and P. Reusch, “How to develop lean project management?,” Proc. 2013 IEEE 7th Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. IDAACS 2013, vol. 2, no. September, pp. 547–550, 2013.

[16] A. N. A. Wahab, M. Mukhtar, and R. Sulaiman, "A Conceptual Model of Lean Manufacturing Dimensions," Procedia Technol., vol. 11, no. Iceei, pp. 1292–1298, 2013.

[17] U. Viswanath, "Lean transformation: How lean helped to achieve quality, cost and schedule: A case study in a multi location product development team," Proc. - 2014 IEEE 9th Int. Conf. Glob. Softw. Eng. ICGSE 2014, pp. 95–99, 2014.

[18] T. Schröders and V. Cruz-machado, "Industrial Engineering, Management Science and Applications 2015," vol. 349, pp. 803–811, 2015.

[19] A. Chiarini, P. Found, and N. Rich, "Understanding the lean enterprise: Strategies, methodologies, and principles for a more responsive organization," Underst. Lean Enterp. Strateg. Methodol. Princ. a More Responsive Organ., pp. 1–287, 2015.

[20] R. A. Perdana, A. Suzianti, and R. Ardi, "Crowdfunding website design with lean product process framework," Proc. 3rd Int. Conf. Commun. Inf. Process. - ICCIP '17, pp. 369–374, 2017.

[21] B. Fitzgerald, M. Musiał, and K.-J. Stol, "Evidence-based decision making in lean software project management," Companion Proc. 36th Int. Conf. Softw. Eng. - ICSE Companion 2014, pp. 93–102, 2014.

[22] J. M. H. Costa, M. Rossi, E. Rebentisch, S. Terzi, M. Taisch, and D. Nightingale, "What to measure for success in Lean system engineering programs?," Procedia Comput. Sci., vol. 28, pp. 789–798, 2014.

[23] E. Kerga, A. Rosso, W. Bessega, A. Bianchi, C. Moretti, and S. Terzi, "Compact Teams: a Model to Achieve Lean in Product Development," 2006.

[24] D. Powell, A. Binder, and E. Arica, "MES support for lean production," IFIP Adv. Inf. Commun. Technol., vol. 398, no. PART 2, pp. 128–135, 2013.

[25] G. Schuh and P. Stüer, "Framework for lean management in industrial services," IFIP Adv. Inf. Commun. Technol., vol. 398, no. PART 2, pp. 392–398, 2013.

[26] B. Resta, P. Gaiardelli, S. Dotti, and R. Pinto, "Towards a New Model Exploring the Effect of the Human Factor in Lean Management," Adv. Prod. Manag. Syst. Innov. Prod. Manag. Towar. Sustain. Growth (Amps 2015), Pt Ii, vol. 460, pp. 316–323, 2015.

[27] I. Butterworth, K. Westwood, B. Hill, W. Midlands, and J. Brighton, "Proceedings of the 11th International Conference on Manufacturing Research (ICMR2013)," no. Ichd, pp. 1–8, 2012.

[28] U. Viswanath, "Lean Transformation," Proc. 9th India Softw. Eng. Conf. - ISEC '16, pp. 156–162, 2016.

Call for Papers/Authors Guideline

KIET Journal of Computing & Information Sciences (KJCIS) is biannual publication of College of Computing & Information Sciences, Karachi Institute of Economics and Technologies. It is published in January and July every year. We are lucky to have on board prominent and scholarly academicians as part of Advisory Committee and reviewers.

KJCIS is a multi-disciplinary journal covering viewpoints/ researches / opinions relevant to the non exhaustive list of the topics including data mining, big data, machine learning, artificial intelligence, mobile applications, computer networks, cryptography & information security, mobile and wireless communication, adhoc & body area networks, software engineering, speech & pattern recognition, evolutionary computation, semantic web & its application, data base technologies & its applications, Internet of Things (IoT), computer vision, distributed computing, grid and cloud computing.

The authors may submit manuscripts abiding to following rules:-

- Certify that the paper is original and is not under consideration for publication in any other journal. Please mention so, in case it has been submitted elsewhere.
- Adhere to normal rules of business or research writing. Font style be 12 points and the length of the paper can vary between 3000 to 5000 words.
- Illustrations/tables or figures should be numbered consecutively in Arabic numerals and should be inserted appropriately within the text.
- The title page of the manuscript should contain the Title, the Name(s), email address and institutional affiliation, an abstract of not more than 200 words should be included. A footnote on the same sheet should give a short profile of the author(s).
- Full reference and /or websites link, should be given in accordance with the APA citation style. These will be listed as separate section at the end of the paper in bibliographic style. References should not exceed 50.
- All manuscripts would be subjected to tests of plagiarism before being peer reviewed.
- All manuscripts go through double blind peer review process .
- Electronic submission would only be accepted at kjcis@pafkiet.edu.pk
- All successful authors will be remunerated adequately.
- The Journal does not have any article processing and publication charges.

Submission is voluntary and all contributors will find a respectable acknowledgment on their opinion and effort from our team of editors. Submission of a paper will be held to imply that it contains original unpublished work. In case the paper has been forwarded for publication elsewhere, kindly apprise in time if the paper has been accepted elsewhere. Manuscripts may be submitted before September and May to get published in Jan & July issues respectively. We encourage you to submit your manuscripts at kjcis@pafkiet.du.pk

Editorial Board KJCIS
College of Computing & Information Sciences
Karachi Institute of Economics and Technology

Karachi Institute of Economics and Technology
Korangi Creek, Karachi-75190, Pakistan
Tel: (9221) 3509114-7, 34532182, 34543280 Fax: (92221) 35009118
Email: kjcis@pafkiet.du.pk
http://kjcis.pafkiet.edu.pk

KARACHI INSTITUTE OF ECONOMICS
AND TECHNOLOGY

PUBLISHED BY:

College of Computing and Information
Sciences, KIET.
www.cocis.pafkiet.edu.pk
www.kjcis.pafkiet.edu.pk
kjcis@pafkiet.edu.pk

MAIN CAMPUS

PAF BASE Korangi Creek,
Karachi-75190.
Ph: (9221) 35091114-7
Fax: (9221) 35091118

CITY CAMPUS

28-D, Block 6, P.E.C.H.S,
Karachi-75400.
Ph: (9221) 34543280
Fax: (9221) 34383819

NORTH NAZIMABAD CAMPUS

F-98, Block B, (Near KDA roundabout)
North Nazimabad, Karachi-74700.
Ph: (9221) 36628381 or 36679314
Cell: 0336-2444191-92

Printed by HAWK