International Journal of Computer Engineering and Technology (IJCET)
Volume 15, Issue 1, Jan-Feb 2024, pp. 7-15, Article ID: IJCET_15_01_002
Available online at https://iaeme.com/Home/issue/IJCET?Volume=15&Issue=1
ISSN Print: 0976-6367 and ISSN Online: 0976–6375
Impact Factor (2024): 18.59 (Based on Google Scholar Citation)
DOI: https://doi.org/10.17605/OSF.IO/KWAFG

ADVANCING PERSONALIZED MEDICINE: A STRATEGIC FRAMEWORK FOR INTEGRATING NON-TRADITIONAL DATA SOURCES

Lakshmi Sahitya Cherukuri
Independent Researcher, Dallas, TX, USA

ABSTRACT

This paper presents a strategic framework for advancing personalized medicine through the integration of traditional and non-traditional data sources. Recognizing the limitations of relying solely on clinical records, genomic data, and clinical trials, a comprehensive approach is proposed that includes data from online social networks and discussion forums, wearable devices, and patient-generated health data. The importance of these diverse sources is emphasized in capturing real-time insights into patient behaviors and responses, which are often missed in standard clinical settings. This approach employs Large Language Models (LLMs) to efficiently integrate and analyze the heterogeneous and unstructured nature of these data streams. This integration facilitates a more holistic view of health management and outcomes, enhancing the precision of personalized medicine. The complexities of combining structured clinical information with unstructured patient narratives and digital interactions are discussed in the paper, highlighting the role of LLM in extracting relevant health information. The paper also delves into the significance of alternative data sources like social media, wearable technology, and sensor data from smart devices in refining predictive analytics and improving patient care. Additionally, challenges and opportunities in data integration are discussed, emphasizing the need for a robust data strategy that encompasses data quality, stewardship, exchange, and analytics. Proposed methodology contributes to the field by offering a data-driven approach for health risk assessment and management, ultimately aiming to improve patient outcomes and healthcare practices in the domain of personalized medicine.

Keywords: Digital Twins, Large Language Model, Non-Traditional Data, Personalized Medicine, Unstructured Data

Cite this Article: Lakshmi Sahitya Cherukuri, Advancing Personalized Medicine: A Strategic Framework for Integrating Non-Traditional Data Sources, International Journal of Computer Engineering and Technology (IJCET), 15(1), 2024, 7-15.
https://iaeme.com/Home/issue/IJCET?Volume=15&Issue=1

1. INTRODUCTION

The emergence of personalized medicine, driven by the analysis of large volumes of health-related data, has marked a significant advancement in healthcare. With new mobile applications, devices and monitoring technologies entering the market every day, there is an explosion of health data. However, this innovative data-driven approach encounters substantial challenges in integrating data from diverse sources. Some of the challenges and techniques to overcome are discussed here [1]. Traditional data sources like clinical records, genomic data, clinical trials, and patient registries, often do not fully capture the dynamic nature of individual health experiences. People are increasingly sharing their well-being, side effects, experiences with new drugs in the connected applications. Integrating non-traditional data sources, such as online forums, wearables, and patient-generated health data, is crucial for a comprehensive view of health management and outcomes. These sources offer better insights into patient behaviors and help providers assess the risk of an intervention. This paper proposes an approach that utilizes advancements in transformer based Deep Learning (DL) architecture model, LLM, to efficiently integrate and analyze these diverse data streams, thereby enhancing the precision of personalized healthcare.

The success of new technologies is dependent on how well tools are used for ongoing research and ability to seamlessly integrate new data with current systems. Siloed approaches for data capture from different organizations following unique set of standards pose challenges [1] when integrating the data. The variety of data types, ranging from structured clinical information to unstructured patient narratives and digital interactions, necessitates sophisticated methods for harmonization and interpretation. Data heterogeneity and unstructured nature present significant challenges in data management and analysis.

Non-traditional data sources are invaluable for providing feedback on health interventions, capturing patient behaviors [2] and experiences that are often overlooked in clinical settings. These sources shed light on how lifestyle choices, environmental factors, and personal habits impact health outcomes, a study is conducted to evaluate health outcomes [3]. To effectively harness this data, the paper explores the use of LLM for extracting relevant health information from unstructured data, aiming to provide a more holistic and accurate assessment of health risks and outcomes.

A literature review on personalized medicine implementation with non-traditional data sources is studied by [4]. Knowledge graphs with multiple nodes or data sources with unique entities, show the lineage and inter dependencies of data points is another avenue for better knowledge inference [5]. Biomedical ontologies contain metadata for medical terminology. With the combination of extracting text using NLP algorithm and biomedical ontology, data extraction from multiple domains is reviewed by Markus et el. [6]. Most existing literature addressed the problem of data integration of structured and unstructured data through other intelligent ways prevalent at the time of research. LLM is a relatively new and developing field of interest.

Machine Learning algorithms, including those based on Natural Language Processing (NLP), have shown enhanced accuracy with a variety of data. This paper introduces a concept for integrating and analyzing diverse health data sources using LLMs as a key solution for converting unstructured data into structured data.

By leveraging LLMs for advanced text extraction and data processing, this research aims to bridge the gap between structured clinical data and unstructured, non-traditional data sources. The utilization of LLMs in this context is expected to facilitate a more comprehensive and closer to accurate assessment of individual health risks and responses to interventions. This paper significantly contributes to the evolution of personalized medicine by proposing a robust methodology that employs LLMs for data-driven health risk assessment and management. This approach is designed to enhance patient outcomes and healthcare practices by efficiently processing and analyzing vast amounts of diverse health data.

2. IMPORTANCE OF NON-TRADITIONAL DATA

Alternative data sources offer a wealth of information that is instrumental in refining predictive analytics, identifying candidates for clinical trials, and pinpointing at-risk populations for targeted interventions. For example, electronic health records (EHRs) can be used to identify patients with specific medical conditions who may be suitable for clinical trials. Similarly, credit card data can reveal consumer spending patterns, providing insights into changes in health status. Wearable devices, a rapidly growing segment in the Internet of Things (IoT) healthcare market [7], offer real-time data on patient activity levels, which is particularly valuable for managing chronic conditions such as diabetes or heart disease.

The market for IoT healthcare devices, such as wearable fitness trackers, sleep diary apps, and smart scales, has seen steady growth. These devices collect, send, and store data autonomously, providing healthcare providers with valuable insights into various health indicators. Applications like CareKit, ResearchKit, and HealthKit facilitate the sharing of IoT user data with healthcare providers and researchers, enabling continuous monitoring and tracking of patient health indicators. This capability is invaluable for early intervention and prevention efforts. For instance, patients with heart failure showing decreased physical activity can be proactively contacted for check-ups, while those with fluctuating blood sugar levels may receive more in-depth nutritional guidance.

Non-medical factors that influence the health outcomes are termed as Social Determinants of Health (SDOH), which has gained significant traction in recent years. Factors such as economic stability, education, social context, neighborhood, and environment play a substantial role in health outcomes and healthcare utilization. Recognizing this, major healthcare organizations have developed programs to address SDOH, with hospitals investing billions in related initiatives [8]. The Centers for Disease Control and Prevention (CDC) defines SDOH as conditions in the environments where people live, learn, work, and play that affect a wide range of health risks and outcomes. According to the National Academy of Medicine, SDOH accounts for a significant portion of the modifiable contributors to health outcomes for a population, underscoring the importance of including these factors in patient profiles for population health and preventive care programs.

Non-traditional data sources, including online social media communities, online forums, patient-generated information from wearables, consumer spending patterns from credit card data, and SDOH, provide a richer, more comprehensive understanding of patient health and behavior. This expanded data landscape enables healthcare providers to make more informed decisions, design targeted interventions, and improve the overall value of healthcare delivery closing the loop of care effectiveness.

The integration of these diverse data sources is supported by a robust data strategy that focuses on data quality, stewardship, exchange, and analytics. Artificial Intelligence (AI) and algorithmic decision-making, including machine learning, play a pivotal role in optimizing the use of this data.

The increasing reliance on patient-reported outcomes and electronic patient-reported outcomes further emphasizes the patient as a central contributor to healthcare data, especially in remote healthcare settings.

An integrated data ecosystem is essential for leveraging these disparate data sources successfully. This ecosystem combines structured and unstructured data, leading to numerous benefits in clinical research and patient care. These benefits include more accurate patient recruitment for clinical trials, faster validation of trial designs, and the identification of data trends to support research endpoints. Researchers gain real-time access to data, which is crucial for fast-track regulatory submissions and informed decision-making across the healthcare landscape.

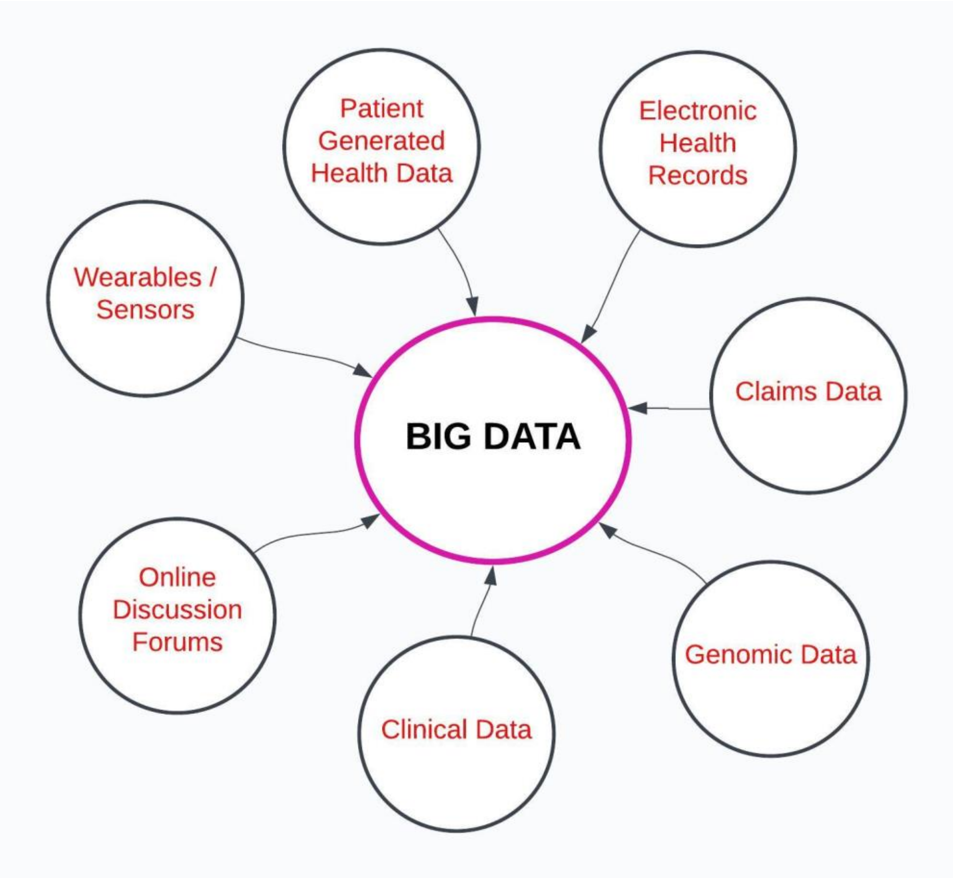

Fig. 1. Non-Traditional Disparate Data Sources in Healthcare ecosystem

3. UNDERSTANDING ALTERNATIVE DATA

3.1. Online Discussion Forums

Online discussion forums and social networks have become vital sources of health-related information [9], where individuals share personal experiences and seek comfort during challenging health situations. Platforms like Reddit, Yahoo! Answers, and MedHelp.com host vibrant communities where users engage in health-related conversations. These discussions often take the form of threads with multiple replies, contributing to a rich tapestry of shared experiences and advice. Research, such as the study published in the Journal of Medical Internet Research [10], has validated the quality of information in these forums, highlighting their usefulness. The data in these forums is typically unstructured, comprising free-text responses that reflect real-world evidence often missed in clinical settings.

This unstructured data is invaluable in personalized medicine for understanding patient experiences, treatment reactions, and the effectiveness of interventions from a patient's perspective.

3.2. Patient-Generated Health Data from Mobile Applications

Mobile health applications are increasingly used to collect patient-generated health data, especially for managing specific diagnoses. These apps, often provided by treatment providers, encourage patients to enter relevant data about their condition, medication effects, and physiological factors. The data collected is usually in a semi-structured format, combining structured data entry (like numerical values for blood pressure) with unstructured data (such as patient descriptions of symptoms or side effects). This data is crucial in personalized medicine for tracking the effectiveness of treatments [11] over time and understanding the patient's experience with their condition.

3.3. Wearable Device Data

Wearable devices, including smartwatches, fitness trackers, and health-monitoring patches, offer a continuous stream of structured health data [12]. They monitor various physiological parameters like heart rate, blood pressure, sleep patterns, and physical activity. This data is collected in a structured format, providing researchers with a comprehensive dataset that captures trends and variability over time. Wearables provide non-invasive insight into physiological signals from the body enabling precision medicine.

3.4. Sensor Data from Smart Devices

Smart devices equipped with sensors provide surveillance data that can be mapped with physical behaviors to assess any visible changes from the treatment interventions. This data is a continuous stream, recorded at regular intervals. Alerts from smart devices act as crucial feedback when an individual is undergoing a treatment, allowing for adjustments in treatment plans based on real-time motion.

3.5. Challenges and Opportunities in Data Integration

The data from these sources comes in multiple file formats and structures, with varying access protocols and semantics, reflecting the diverse intentions of the platforms collecting the data. Integrating this multimodal data requires careful consideration of its structure and semantics. Data storage, data analysis and the technologies required to operate on structured and unstructured data are different. Each problem of type of assessment, text retrieval, inferencing on the sentiment of the experience demands different approaches. For personalization support, this integration is essential for a holistic understanding of patient health, encompassing clinical data, lifestyle factors, and patient-reported outcomes. The effective use of this diverse data can lead to more accurate diagnoses, tailored treatment plans, and improved patient outcomes, marking a significant step forward in the field of personalized healthcare.

4. DIFFERENT CATEGORIES OF DATA

4.1. Structured Data

Structured data is characterized by its clear definition and organization within a relational framework. This type of data typically includes formats that are easily searchable and storable in traditional database systems, such as Relational Databases. Commonly sourced from online relational and tabular forms, structured data is ideal for machine learning applications due to its high level of organization and predefined schemas.

It is stored in data warehouses, which are optimized for handling structured data, ensuring efficient querying and analysis. The structured nature of this data means it is not as flexible as other forms, but it requires less storage space compared to unstructured or semi-structured data. Examples of structured data include tables in relational database management system (RDBMS), where each column and row represent a specific type and value of data, respectively.

4.2. Unstructured Data

Unstructured data, in contrast, lacks a predefined format or organization, making it more complex to process and analyze. This data type encompasses a wide range of formats, including images, videos, emails, documents, social media content, and more. It is particularly suited for natural language processing and text mining, as these techniques are designed to extract meaningful information from data that does not fit into traditional relational databases. Unstructured data is typically stored in data lakes, Hadoop, which are designed to store vast amounts of raw data in its native format. The flexibility of unstructured data allows for a broader range of information to be captured, but it also requires significantly more storage space. Examples of unstructured data include various file types like JPEG images, DOC files, PDF documents, and MOV video files, each containing a wealth of information that is not immediately accessible in a structured format.

4.3. Semi-Structured Data

Semi-structured data strikes a balance between the rigid organization of structured data and the free-form nature of unstructured data. It is semi-defined, often containing tags or markers to separate semantic elements and enabling the grouping of similar data. This type of data is commonly used in natural language processing and text mining, like unstructured data, but with the added advantage of some level of organization. Semi-structured data is often sourced from web documents, JSON, and XML files, which contain both structured and unstructured elements. It can be stored in both data warehouses and data lakes, benefiting from the features of both storage systems. Semi-structured data is somewhat flexible, offering more adaptability than strictly structured data, and typically requires a medium amount of storage space. Examples include JSON, XML, CSV files, and databases that can store that type of data are MongoDB, Cassandra.

5. LARGE LANGUAGE MODELS FOR DATA EXTRACTION

Large Language Models like GPT-3, GPT-4, and the LLaMA series, built on deep learning architectures including BERT and transformers, have revolutionized Natural Language Processing. Trained on extensive datasets, these models excel in retaining large contexts and performing various NLP tasks. They operate through complex neural networks, where each node contributes to the final output, finely tuned to derive parameters that transform input data into meaningful results. Smaller LLMs, with fewer weights, offer the advantage of a reduced memory footprint, resulting in faster and more cost-effective operations. Crucially, LLMs are adept at text extraction, a key NLP task involving the identification and retrieval of specific information from unstructured data. This includes Named Entity Recognition (NER), Information Retrieval, Document Summarization, Data Categorization, Pattern Recognition, Sentiment Analysis, and Custom Extraction Tasks. The effectiveness of LLMs in these areas hinges on the quality of training data and may require task-specific fine-tuning for optimal performance.

Structured data is well-organized and easy to analyze, often stored in relational databases where Structured Query Language (SQL) is used to manage and query the data. For handling large amounts of data, data lakes and data warehouses are common storage options. Incorporating Large Language Models significantly enhances our capability to analyze multimodal data, which combines various unstructured data types. LLMs are adept at interpreting and generating human-like text, making them invaluable in processing and extracting information from large text corpora using NLP techniques. This integration enables the extraction of nuanced insights from multimodal data sources, where LLMs can provide context and augment the interpretation of image and textual data. As a result, LLMs play a crucial role in bridging the gap between diverse data types, offering a comprehensive framework for understanding and leveraging the wealth of information contained within multimodal unstructured data.

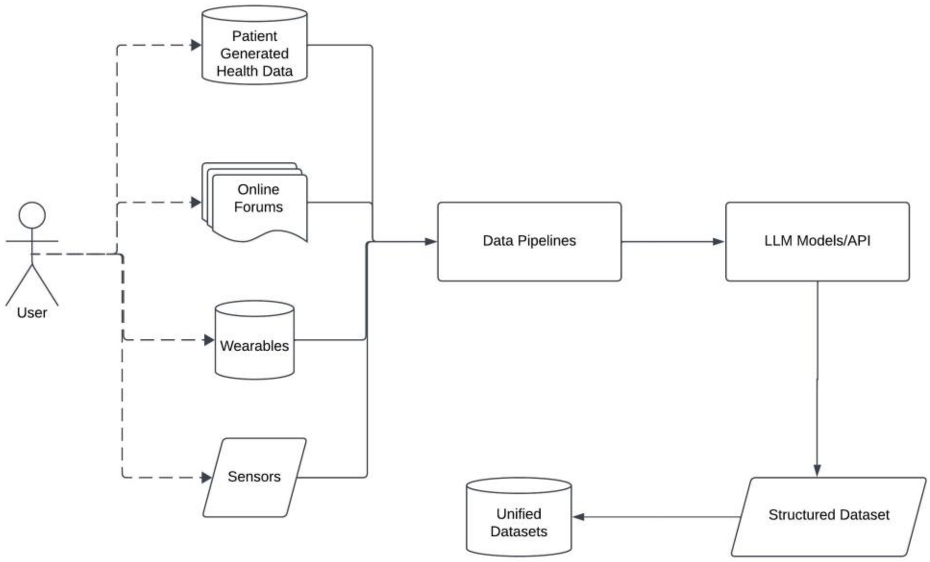

Fig. 2. Dataset conversion from unstructured to structured

6. LIMITATIONS AND FUTURE WORK

In this research, several key challenges and future considerations were identified when using LLMs for converting unstructured data to structured data. Firstly, understanding how these models work internally is complex. There's a whole specialized field, known as mechanistic interpretability, dedicated to dissecting the technical intricacies of transformer neural network architecture. Choosing the right LLM model is crucial, it depends on the size of the data and the specific context needed for answering questions related to the topic. Training these models can be costly, especially when considering the number of model parameters, GPU requirements, and organization's budget constraints. When using open-source or custom models, fine-tuning may be necessary to enhance accuracy and tailor them to specific use cases. While there are studies exploring the use of unstructured data in its original form [13], any future advancements will demand significant resources and collaboration among various stakeholders. Although LLMs are adept at understanding context, they may struggle with domain-specific nuances if not pre-trained in those areas. The process of transforming large volumes of unstructured data into a structured format is resource-intensive and expensive. Additionally, using LLMs raises concerns about the privacy and security of sensitive information in the data. Lastly, while progress is being made, LLMs are currently more effective with text data compared to non-textual sources like images or videos. These limitations and future research directions are crucial for the continued advancement and application of LLMs in data processing.

REFERENCES

[1] Martínez-García M, Hernández-Lemus E. Data Integration Challenges for Machine Learning in Precision Medicine. Front Med (Lausanne). 2022 Jan 25;8:784455. doi: 10.3389/fmed.2021.784455. PMID: 35145977; PMCID: PMC8821900.

[2] Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating Patient-Generated Health Data Into Clinical Care Settings or Clinical Decision-Making: Lessons Learned From Project HealthDesign. JMIR Hum Factors. 2016 Oct 19;3(2):e26. doi: 10.2196/humanfactors.5919. PMID: 27760726; PMCID: PMC5093296.

[3] Li L, Novillo-Ortiz D, Azzopardi-Muscat N, Kostkova P. Digital Data Sources and Their Impact on People's Health: A Systematic Review of Systematic Reviews. Front Public Health. 2021 May 5;9:645260. doi: 10.3389/fpubh.2021.645260. PMID: 34026711; PMCID: PMC8131671.

[4] Taylor CO, Tarczy-Hornoch P. Personalized Medicine Implementation with Non-traditional Data Sources: A Conceptual Framework and Survey of the Literature. Yearb Med Inform. 2019 Aug;28(1):181-189. doi: 10.1055/s-0039-1677916. Epub 2019 Aug 16. PMID: 31419830; PMCID: PMC6697507.

[5] Abu-Salih B, Al-Qurishi M, Alweshah M, Al-Smadi M, Alfayez R, Saadeh H. Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities. J Big Data. 2023;10(1):81. doi: 10.1186/s40537-023-00774-9. Epub 2023 May 28. PMID: 37274445; PMCID: PMC10225120.

[6] Kreuzthaler M, Brochhausen M, Zayas C, Blobel B, Schulz S. Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems. Front Med (Lausanne). 2023 Mar 15;10:1073313. doi: 10.3389/fmed.2023.1073313. PMID: 37007792; PMCID: PMC10050682.

[7] Khan, S., Alam, M. Wearable Internet of Things for Personalized Healthcare: Study of Trends and Latent Research. In: Patgiri, R., Biswas, A., Roy, P. (eds) Health Informatics: A Computational Perspective in Healthcare. Studies in Computational Intelligence, (Singapore: Springer, 2021) vol 932

[8] Horwitz L, Chang C, Arcilla H, Knickman J. Quantifying Health Systems’ Investment In Social Determinants Of Health, By Sector, 2017–19. Health Affairs 2020 39:2, 192-198. doi: 10.1377/hlthaff.2019.01246

[9] Tanis M. Health-Related On-Line Forums: What's the Big Attraction?, Journal of Health Communication, 2008 13:7, 698-714, doi: 10.1080/10810730802415316

[10] Cole J, Watkins C, Kleine D. Health Advice from Internet Discussion Forums: How Bad Is Dangerous? J Med Internet Res 2016;18(1):e4. doi: 10.2196/jmir.5051

[11] Nittas V, Lun P, Ehrler F, Puhan MA, Mütsch M. Electronic Patient-Generated Health Data to Facilitate Disease Prevention and Health Promotion: Scoping Review. J Med Internet Res. 2019 Oct 14;21(10):e13320. doi: 10.2196/13320. PMID: 31613225; PMCID: PMC6914107.

[12] Deng Z, Guo L, Chen X, Wu W. Smart Wearable Systems for Health Monitoring. Sensors (Basel). 2023 Feb 23;23(5):2479. doi: 10.3390/s23052479. PMID: 36904682; PMCID: PMC10007426.

[13] Luo L, Li L, Hu J, Wang X, Hou B, Zhang T, Zhao LP. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC Med Inform Decis Mak. 2016 Aug 30;16(1):114. doi: 10.1186/s12911-016-0357-5. PMID: 27577240; PMCID: PMC5006527.

Citation: Lakshmi Sahitya Cherukuri, Advancing Personalized Medicine: A Strategic Framework for Integrating Non-Traditional Data Sources, International Journal of Computer Engineering and Technology (IJCET), 15(1), 2024, 7-15.

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJCET/VOLUME_15_ISSUE_1/IJCET_15_01_002.pdf

Abstract Link:
https://iaeme.com/Home/article_id/IJCET_15_01_002

editor@iaeme.com