ROLE OF DATA MINING IN MEDICAL HEALTHCARE

By

WAQAS ALI

NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

October 2022

ROLE OF DATA MINING IN MEDICAL HEALTHCARE

By

WAQAS ALI

MCS, ARID University, Rawalpindi, 2018

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In

Software Engineering

To

FACULTY OF ENGINEERING & COMPUTER SCIENCES

NATIONAL UNIVERSITY OF MODERN LANGUAGES ISLAMABAD

THESIS DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defense, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Software Engineering.

THESIS TITLE: Role of Data mining in Medical Healthcare.

Waqas Ali	14 MSSE/IBD/S-19
Submitted by	Registration #
Master in Software Engineering	Software Engineering
Title of the degree	Name of Discipline
Dr. Raheel Zafar	_____
Name of the Research Supervisor	Signature of the Research Supervisor
Dr. Muhammad Javvad ur Rehman	_____
Name of Co-supervisor	Signature of Co-supervisor
Dr. Basit Shahzad	_____
Name of Dean (FE&SE)	Signature of Dean (FE&CS)
Brig Syed Nadir Ali	_____
Name of Director General	Signature of Director General

OCTOBER 10TH 2022
Date

"I hereby declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality for the award of the degree of Masters of Science in *(Software Engineering)*"

Signature : _____

Name : Dr. Raheel Zafar

Date : 10th October 2022

Signature : _____

Name : Dr. Muhammad Javvad ur Rehman

Date : 10th October 2022

AUTHOR'S DECLARATION

I Waqas Ali

Son of Ghulam Ali Malik

Registration # 14-MSSE/IBD/S19

Discipline Software Engineering

Candidate of **Master of Science in Software Engineering (MSSE)** at the National University of Modern Languages do hereby declare that the thesis **Role of Data Mining in Medical Healthcare** submitted by me in partial fulfillment of MSSE degree, is my original work, and has not been submitted or published earlier. I also solemnly declare that it shall not, in future, be submitted by me for obtaining any other degree from this or any other university or institution. I also understand that if evidence of plagiarism is found in my thesis/dissertation at any stage, even after the award of a degree, the work may be cancelled and the degree revoked.

Signature of Candidate

Waqas Ali
Name of Candidate

10th October 2022
Date

ABSTRACT

ROLE OF DATA MINING IN MEDICAL HEALTHCARE

Data mining (DM) is a progressive field that helps in finding useful and meaningful information from large data. It aids to determine knowledge and patterns from complex data. Health data needs various investigative procedures in identifying vital information that is used for decision-making. In healthcare, organization's data is mostly stored in digital format all over the world. Enhancement is always an important feature to examine. In the medical healthcare field, various researchers are interested to contribute accordingly. The data of medical healthcare exists but it needs more attention by applying DM techniques and sort in a more compatible form of knowledge. However, the lack of a comprehensive and systematic narrative encourages for bearing a systematic literature review (SLR) on this topic. This research aims to find updated knowledge of DM and machine learning (ML) techniques in medical healthcare. The comparison is prepared based on three different methods which include quantitative-based, image-based, and signals-based. In this study, SLRs focus on the published literature of a specific research field by the findings of all relevant studies that address a set of research questions while being objective, systematic, clear, and replicable. Firstly, this study answers the current status of DM, ML, and their algorithms. Secondly, three different methods are compared and propose a framework to help the data analysts and data science experts to know about the suitable DM, ML techniques, and methods for medical healthcare.

TABLE OF CONTENTS

LIST OF TABLES

S.NO **TITLE** **PAGE**

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	-	Alzheimer Disease
ANN	-	Artificial Neural Network
AUC	-	Area under the Curve
BDA	-	Big Data Analytics
CNN	-	Convolutional Neural Networks
COVID-19	-	Corona Vires Identity 2019
CKD	-	Chronic Kidney Disease
CT	-	Computed Tomography
DM	-	Data Mining
DL	-	Deep Learning
DS	-	Data Science
DT	-	Decision Tree
ECG	-	Electrocardiogram
EEG	-	Electroencephalogram
FMRI	-	Functional Magnetic Resonance Imaging
IEEE	-	Institute of Electrical and Electronics Engineers
KNN	-	k-Nearest Neighbors
LR	-	Linear Regression
LR	-	Logistic Regression
MDPI	-	Multidisciplinary Digital Publishing Institute
ML	-	Machine Learning
MRI	-	Magnetic Resonance Imaging
NB	-	Naive Bayes
QA	-	Quality Assessment
RF	-	Random Forest
SLR	-	Systematic Literature Review
SVM	-	Support Vector Machine
WEKA	-	Waikato Environment for Knowledge Analysis
X-RAY	-	X-Radiation
XG boost	-	eXtreme Gradient Boosting

ACKNOWLEDGEMENT

First and foremost, I want to convey my gratitude to Almighty Allah, who made this study feasible and fruitful. This research would not have been possible without the sincere support provided by several sources, for which I am grateful. However, there were major contributors to my achievement, and I will never forget their contributions, particularly my research supervisors, Dr. Raheel Zafar and Dr. Muhammad Javvad ur Rehman, who did not leave any stone unturned in their efforts to help me during my research journey.

I like to thank the department of software engineering's administration for their continued support and help during my research experience, which made the hurdles I faced much easier to overcome. Thank you to everyone who I didn't include but who made a big contribution that I will not overlook.

DEDICATION

This thesis is dedicated to my parents and instructors throughout my educational career, who have not only unconditionally loved me but also encouraged me to work hard for the things I want to achieve.

CHAPTER 1

INTRODUCTION

1.1 Overview

In the first chapter, the present state of data mining (DM) research in healthcare is discussed. To begin, the entire context of the planned investigation is outlined. Existing literature is used to identify the research problem. Keep in mind the notion of solutions to a study topic. The research topics that develop as a result of this study are organized according to SLR principles. The goals are also given to meet the needs of value additions. This study tries to provide a clear picture of the research's direction. The breadth of the study is vital, as is how it will be valuable in the medical field and how it will be implemented in current literature structure. After that, significant findings from this research will be discussed, and these discoveries will be added to the current body of knowledge. The thesis outlines are given in full details, at the end of this chapter, for the convenience of readers.

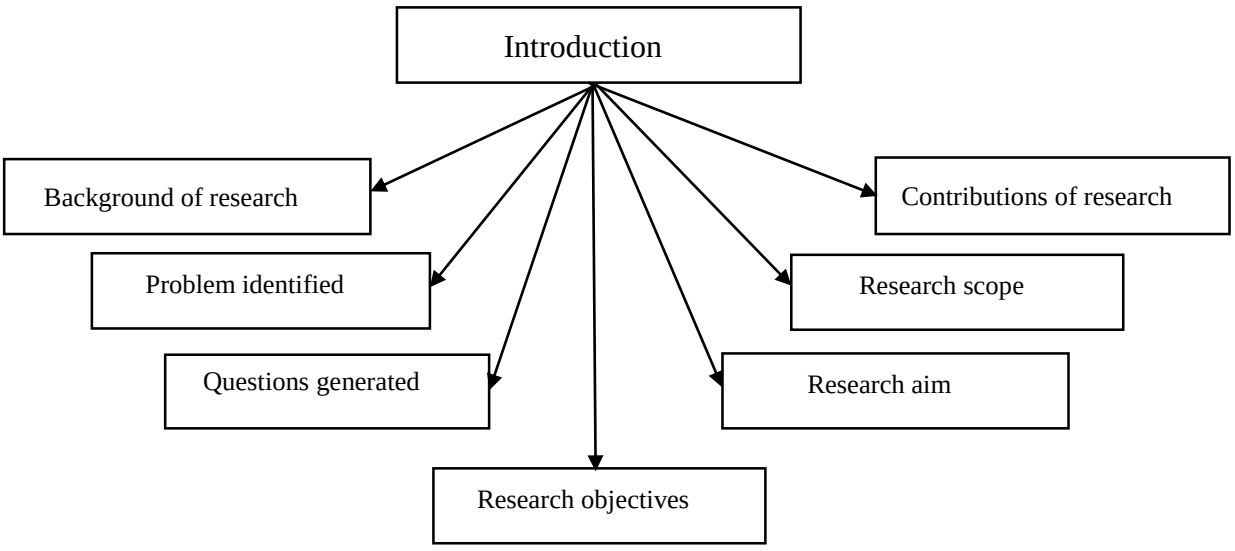

Figure 1.1: Layout of the chapter 1

1.2 Background of research

The process of extracting useful information from a huge amount of data is called DM. It is a tool used by humans to discover new, accurate, useful patterns in data, and meaningful relevant

information for the ones who need it. DM facilitates getting information from the challenging formation of data. The process of knowledge is performed at this stage. DM is the best proceeding ground of knowledge, which is generating meaningful and helpful information from the bulk of data. The logical methodology is used in health data because of making policies information identified by DM. All over the world digital format is used as data storage for healthcare organizations. The big data of all patient details are contained in that format. The complexity and complications increase due to huge data. This creates problems while using traditional methods to collect information. To make it easy the advancement in mathematics, computer and statistics are used. These fields are putting efforts in healthcare treatments. The healthcare industry require database system that keeps proper identified data regarding treatment of patients. The complex and raw form of data converted into meaningful knowledge is extracted by DM as it is probing of available datasets in order to identify patterns and anomalies [1].

The developing healthcare organizations are producing a large body of appreciated records on patient bio-data, treatment strategies, and expenditures appealing to the dedication of clinicians and experts alike [2]. The healthcare industry requires DM in the revolution of awareness and discovery preparations for policymaking. DM is an extremely evolving scope, which need outcomes that are convenient and useful details from big data. Health requirements are analytical methodology in detecting dynamic material that helps to make perfect policies [3]. Healthcare scopes are searching for suitable skills to justify resources for increasing patient awareness and administrative management. The quality of numerous features of healthcare is enhanced by the predictive, descriptive, and analytical applications [4]. A large size of data is composed and resolved in this system frequently. The knowledge of healthcare is arranged by the analytic tools and techniques which collect information from difficult, huge data and it is transformed for making policies [2].

The goal of DM is to extract the rules from massive amounts of data, whereas the goal of ML is to train a computer to understand and learn the parameters. DM is merely a method of conducting research to identify a specific result based on the sum of data acquired. ML leverages collected data and experience to make systems smarter while training them to execute difficult tasks. To produce forecasts for businesses and other organizations, DM relies on enormous data stores (such as Big Data). On the other hand, ML utilizes algorithms rather than unprocessed data. Additionally, DM is a method that combines the database with ML. While the latter offers data analysis techniques, the

former offers data management techniques. So while DM needs ML, ML doesn't necessarily need DM [1].

For data management and data analysis approaches, DM uses two components: databases and ML. It assists in the extraction of priceless data that might offer outstanding insights about a good or service. However, ML merely employs algorithms and has the power to adapt its rules based on a given scenario to get the best possible outcome. The amount of human work is another stark contrast, with DM requiring continuous human involvement while ML just needs humans to define the algorithm. Compared to DM, ML will work independently to give accurate findings because it is an automated process. DM is restricted to the manner in which data is gathered and organized, and it serves as a tool for drawing out pertinent information from large datasets. ML identifies the correlations between all relevant data points to deliver accurate conclusions and ultimately shape the model's behaviour [3].

The process of discovering algorithms that have improved courtesy of experience-derived data is known as ML. It is the algorithm that permits the machine to learn without human intervention. It's a tool to make machines smarter, eliminating the human element. A large quantity of data produced in organizations by various devices with difficult conditions has importance for examination via ML algorithms. These algorithms are certified to extract valued figures from the developed data and induct useful corollaries. In an accurate environment, ML methods provide great accuracy. Many up-to-date research efforts are pointed at discovering new extents of ML applications to healthcare organizations, evaluating their correctness for systems, and the accuracy rises by prediction and analysis models [5]. The complete processes are finding, treatment, and disease avoidance covers in healthcare. The health industry in many states is developing at a quick pace. It produces massive volume of data, containing digital medical histories, managerial tests, and additional conclusions. The latest improvements in ML tools provide new operative paradigms to get end-to-end knowledge and representations from difficult data [6].

Healthcare invention is the world's most prominent, highly critical, and quick-increasing industry that is developing through significant challenges in the modern era [1]. ML organizations have many advantages, experts use huge capacities of data, named as data for training and complete inductive interpretation, assist the clinical repetition in determining issues and planning remedies. These systems reduce faults by removing manpower essentials and accomplishing repetitive jobs,

therefore improving competence related to physical efforts. Physicians take assistance from artificial intelligence (AI) that knowledge associated with health science from books, published articles, and clinical skills to access suitable patient care. However, the suggestions are identified existing techniques still missing by humans. Observing, handling, and analyzing healthcare reports become easy with the addition of ML in the internet of things (IoT) strategies [5]. When the traditional approaches fail to treat that type of information, the mathematical, statistical, computational and healthcare fields are trying to discover new methods for demonstrating the prediction and identification of human health diseases [3].

The lack of a solution is healthcare production collects vast quantities of medical records that are mined but in a particular data type which is not enough to determine features for reliable policy-making [7]. Previously, scholars attempted to focus their efforts in a single direction. A small number of people are interested in comparing quantitative, image, and signal-based methods using various techniques such as x- radiations (x-rays), computed tomography (CT), magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electrocardiogram (ECG), electroencephalogram (EEG), tabular facts, and figures. Medical care is the most significant topic in human civilizations; citizen's lives are directly dependent on it. It is nevertheless quite diverse, dispersed, and fractured. Assigning appropriate patient care from a systematic standpoint entails gaining access to patient information which is rarely available when and where it is needed [3].

Quantitative data is generated by doctor opinions and recorded in tabular or document form. Like document mining is named text mining, it is used in the medical field to find data on protein. Image data is progressively employed as an ideal diagnostic tool. MRI, CT, x-ray, and fMRI are the instances of image-based methods collected from machines. For example, the health field images have previously been utilized for tumor categorization in digital formats. Signal-based methods are also collected by electronic machines which include ECG and EEG. The graph of ECG in healthcare and the sensor data of EEG is an example of signal-based methods [8].

The SLR was piloted and completely accomplished to collect significant studies. The above-defined methods are compared to estimate the research field. SLR is a set of procedures for selecting the existing material. It is attended by the already uploaded material of a particular emerging area by recognizing, estimating, and participating in the judgments of related involvements. It discovers

established research questions that are objective, systematic, apparent, and replicable. [9]. It is to be discussed in detail later on in the third chapter.

DM used in various domains is important to address data in shape of text, opinion, image, web, and graph of medical data organizations. It has proved to be a significant medical research field for the discovery of unknown outlines in medical data. The healthcare experts can observe the disease estimation inquiry given by the prediction. DM technique shows a dynamic role to predict several diseases. In various cases, doctors cannot guess, a patient is facing different diseases in the meantime. With the beginning of new expansions in the health field, big data about many diseases has been gathered and it is manageable for the research experts [10].

Medical findings in the field of devices and machines are donated as a vital data source for health data. The patient's health-related information, behaviour, physiological parameters, and patient conditions are observed in different devices and sensors to disclose the exact information. Hence, these constant, different, unstructured health-related data resolutions are necessary to accomplish and analyze [11].

Two research questions are discussed in this study. First, what is current status by using applications of DM in healthcare? In the medical health care field, many researchers have contributed to the best of their efforts which is very helpful to evaluate and compare for further investigations. This question is also important to check the current status of medical health care. Secondly, how the comparison is effective among various ML techniques and suggested a framework based on methods in healthcare? The importance of research questions is to collect the latest knowledge related to the DM field. What is the useful contribution of researchers in past decade? Now further putting more efforts into the body of knowledge by comparison among three huge emerging methods.

1.3 Research problem

Several peer-reviewed papers in the medical sector have discussed various aspects of data applications in recent years. Enhancement is a characteristic that should constantly be investigated. Numerous researchers contribute to the medical healthcare profession based on their areas of interest. However, there is still a gap in comparison of identified approaches [4]. Therefore a full comparison is necessary along with the development of a framework to undertake SLR in this area. For increased

comfort in the health sector, multiple methodologies such as quantitative-based, image-based, and signal-based information must be compared. As a result, performance must be improved by incorporating ML algorithms and comparing the three major ways.

1.4 Research questions

RQ1. What is the current status of medical healthcare by using data mining?

RQ2. How the comparison is successful among quantitative, image, and signal-based methods of medical healthcare?

1.5 Research objectives

- To identify current advances in the medical healthcare by using data mining.
- To compare the scoping quantitative, image and signal-based methods of medical healthcare.

1.6 Aim of the research

The goal of this study is to use an SLR to locate useful knowledge. The major focus of this research is on the use of DM approaches and ML algorithms in medical healthcare. In this comprehensive research, three types of methods are employed, including quantitative-based method analysis, image-based method analysis, and how to assess signals-based methods.

1.7 Scope of the research

The SLR guides to appropriate DM approaches and ML algorithms. The goal is to analyse medical applications as a whole, including data collection, pre-processing, artifact removal, feature extractions, feature choices, classifications, correctness, and efficiency. The comparison of quantitative, image, and signals-based methodologies is also highlighted as a beneficial, efficient, and more accurate way for medical treatment. Furthermore, a framework for comparing different methodologies is given, which might provide superior outcomes in the medical healthcare industry.

1.8 Contribution of the research

DM techniques assist in predicting the outcomes or developing a new solution from the past data. By solving DM's issues, ML enables it to expand much more quickly. Additionally, ML is more precise and less prone to errors, giving it the ability to decide for itself and solve problems. The DM process must be maintained, nevertheless, since it will help identifying the issue faced by a certain firm. In order to run a firm and collaborate more effectively, DM and ML are needed. The current advancements in medical healthcare are highlighted in this study. The most up-to-date and successful DM approaches are based on the most recent research. Many approaches that can aid in the medical area are explored. Three quantitative, image-based, and signal-based approaches are used to compare the results. The proposed approach is appropriate for further extensive investigation. This research will assist data analysts and data science specialists in determining the most appropriate DM techniques, ML algorithms, and procedures for medical organizations data as represented by systematic literature knowledge.

1.9 Thesis outline

The first chapter of this proposed study is devoted to the research introduction. The second chapter is a review of existing research or related work. In the third chapter, the methodology employed in this study is described. The fourth chapter explains the analyses and outcomes. Conclusions and future work are discussed in the last chapter. Figure 1.2 represents the whole thesis arrangement which includes all chapter parts like sections and sub-sections.

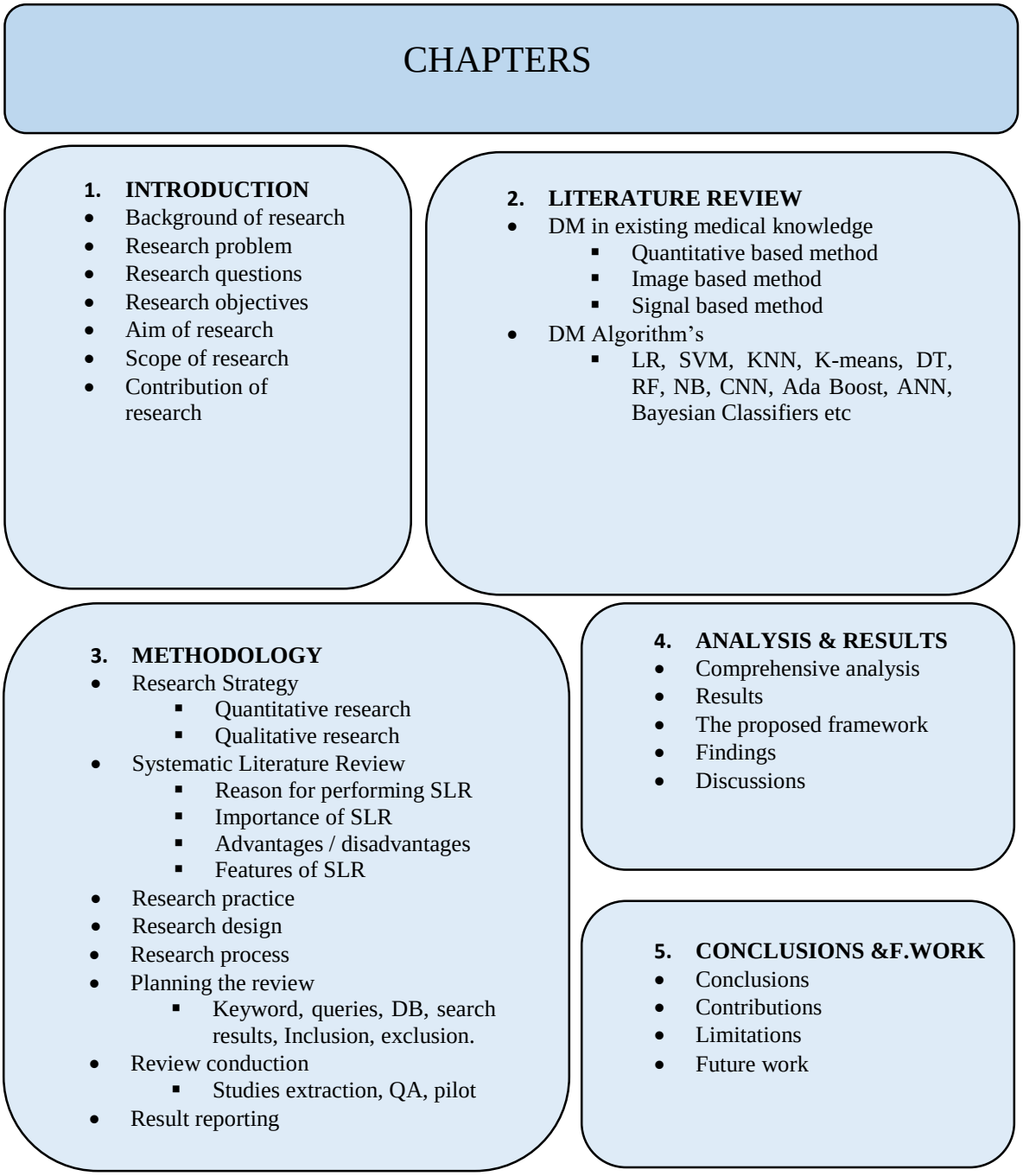

Figure 1.2: Layout of the whole thesis

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Different researches in the medical healthcare area is covered in this chapter, which includes a review of the literature. DM approaches and ML algorithms are used to conclude various research in medical healthcare in this part. The medical field's knowledge is combined and placed on a unified system.

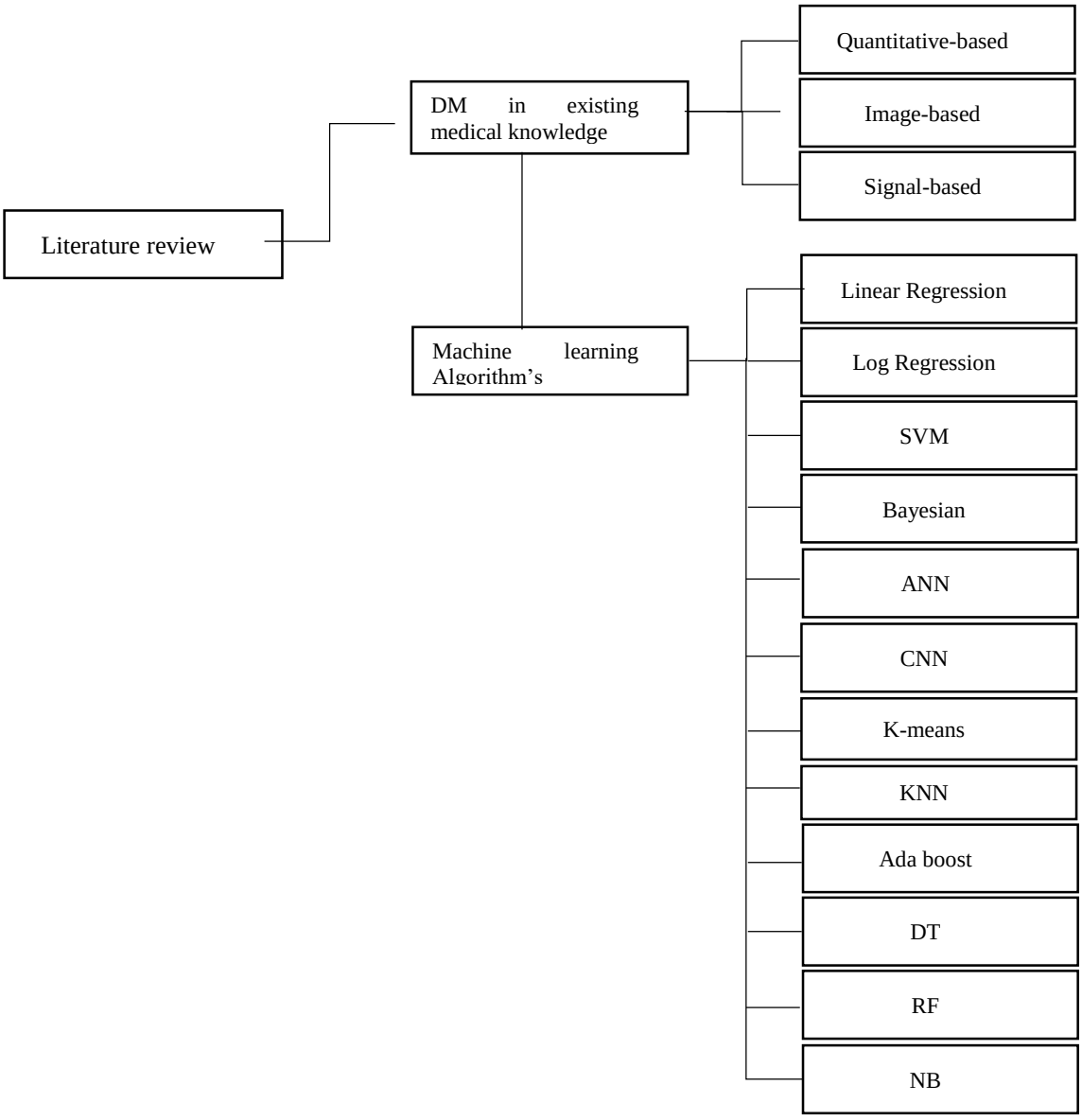

Figure 2.1: Layout of related work

The comparison of three approaches is considered a problem in this study. The first is a quantitative technique, followed by an image-based method, and finally a signal-based way. A tabular representation of the quantitative technique is provided. This sort of data can be gathered from any doctor, social media source, hospital, world health organization, or other relevant agency, for example. The image-based method uses a machine to collect data in the form of X-rays, MRIs, and other imaging techniques.

ML algorithms play an important role in human medical healthcare analysis and are employed in all forms of data analysis. For improved outcomes, all three strategies incorporated ML algorithms. Decision tree DT, logistic regression LR, artificial neural network ANN, k-nearest neighbors KNN, naive bayas NB, convolutional neural network CNN, random forest RF, support vector machine SVM, and others are examples of ML techniques. Table 2.1 summarises previous research by author, title, contributions, technique utilized, benefits, limitations, and publication year.

2.2 Data mining in existing medical knowledge

DM estimates undefined important facts from the bulk of data. DM in healthcare has several assistances are delivered by its users such as the discovery abuse of drug, patients appropriate detection, treatments, early exposure to infections, and survivability [1][12]. The increasing healthcare production is constructing a bulk of valuable facts on patient bio-data, usage tactics, expense, and protection attention, interesting the consideration of clinicians and experts. In the last decades, several articles have talked about different measurements of DM applications in healthcare [2][13][14]. A systematic assessment allocating by the ML is used to the identification of common infections. The existing review attention to new methods associated with the advance in ML. It is used to diagnose patient infections on the health ground, determine motivating outlines, build major guesses and be suitable for making the policies. The results can support investigators to determine and if required to apply in specific areas of ML [3].

The few main subject namely stakeholders, health alertness, organization of hospitals, medical conditions, and healthcare facilities' uses of technology are a demonstration of large data examination in healthcare [4][15]. In fast automation in the healthcare area, IoT is performing an energetic role. The commitment to medical knowledge healthcare (H-IoT) is the subdivision of IoT. The collection

of data and processing is the main element of H-IoT are data elaborate due to big volume in healthcare and exact guesses of the huge value and the combination of ML algorithms into H-IoT is firm [5].

DM is more operational, it creates prediction or clustering demonstrations. There are lots of judgments on both pre-processing complex data and field knowledge expertise. The most current developments in ML knowledge deliver new operational patterns are find end-to-end ML models from challenging data. The researchers review the composed mechanisms for applying ML knowledge to the enhancement of the healthcare industry [6][16]. A DM tool for review of standard open-source in which collection norms established on healthcare presentation desires are scheduled and the best-recommended selection standards are known in the previous study. The standard DM tools of open-source are RapidMiner, KNIME (Which is for incorporating, converting, examining, and arranges of data as a java end-to-end research tool), and R (which is a software development language used for the situation of arithmetic calculating and work of graphics). The existing study displays that RapidMiner and KNIME deliver the largest exposure requirement of DM in healthcare [8].

The application of DM, ML, data analytics, and large data of healthcare industrialised systems were evaluated, combined with published literature [9]. The different infections are predicted by ML algorithms and classification methods i.e. NB, RF, multi-layer perceptron, vote, and sequential minimal optimization (SMO) performance analyzed. It is for multiple data sets of sicknesses i.e. liver, heart disease, diabetes, and chronic kidney disease (CKD). The WEKA is a tool used for disease data from UC Irvine respiratory saved several algorithms was performance evaluated for the experimental setup. Values of many parameters i.e. properly classified occurrences, accuracy, recollection, and f-measure were reserved for analysis by different classification algorithms applications [10]. The different type of features, methods of extracting features, and ML algorithms occupy for electronic Health analysis of data inspect existing studies. The comparison of the best performance for judging healthcare data among neural network (NN), SVM, and other ML algorithms was studied in the literature [11].

The primary studies avail information for systematic reviews to covered every aspect of features [17][18][19]. In the past, the healthcare field data documented in the digital form have constant development in progress but exciting of vast data forced us to think about its preparations into meaningful information. So there is now occurs plenty of knowledge arranged for inquiry.

Investigators are continuously involved in their best struggle to discover an appreciated vision for the excellent facilities in huge data of healthcare [20][21]. Human lives are powerfully influenced by ML systems [22]. The existing study imparts more applications related to DM and also focuses scope of the DM techniques which will be helpful in further research [23][24][25][26][27].

2.2.1 Quantitative-based research

Data is all around world, and every day it becomes increasingly important. Different types of data define more and more of our interactions with the world from using the internet, to increase medical facilities, to the algorithms behind news feeds, and much more. One of the most common and well-known categories of data is quantitative data, or data that can be expressed in numbers or numerical values. Quantitative-based research methods occur in a text or tabular form and it is advised by the doctors. It is purely diagnosed by humans in their own experience. An effective ML-based analysis method was established the finding cardiac sickness. ML algorithms are contained LR, KNN, ANN, SVM, NB, and DT for constructing the method [28]. DM techniques recover the accuracy of estimating cardiovascular sickness. The prediction system was established by multiple groupings of features and techniques of seven classifiers NN, SVM, LR, NB, DT, K-NN, and Vote (a combination of NB and LR). Experiment outcomes indicated that heart sickness achieves 87.4% accuracy by the best performing DM technique (i.e. Vote) [29].

Commonly chest pain complaints were identified in the emergency section of hospitals, and a pilot survey was accomplished in the existing study. A sensitivity of 0.948 and a specificity of 0.546 was achieved by the full model. ML algorithm constructed five main biomarkers were obtained high performance [30]. These existing study findings and predictions several technical studies on medical bases. It was dedicated to the latest research performed by DM techniques to improve the disease predicting process. The author delivers future developments for the latest techniques of knowledge data discovery, by DM tools for the medical field. It was also discussed important issues, DM-associated challenges, and general healthcare. Drug addiction organizations conduct separate interviews of each patient drug abuser [31]. The development of healthcare big data procedure is involved the topic modelling. The big data is accessible by the exposed methods of medical services from medical insurance assessment, facility of evaluation, and their appearance [32].

Big data skill has different areas of presentation in the medical field, like analytical modelling, clinical resolution provision, infection or safety opinion, public health, and investigation. Big data analysis regularly deeds analytic systems established in DM, containing classification, regression, and clustering [33]. The chronic pandemic inflammatory infection is identified as rheumatoid arthritis (RA). The existing study concludes with expert opinions to build various classifiers via some DM techniques. These are to examine the multi-prediction of dual patient groups. Clinical data was possessed for the building of multiple classifiers and to estimate the correctness rate of every classifier subsequently. The best estimate model is nominated from classifiers to forecast the projection of RA within recovery plans and analyze frequent outcomes [34].

Chronic Kidney disease (CKD) managed by traditional Chinese medicine (TCM) is a usage principle and co-prescription design as a capability model of DM. The patient data procedures with CKD were got from the outdoor patient coordination in hospitals of a TCM. The occurrence of a single herb was recommended, as well as their possessions, sense of taste, meridian tropisms, types, also estimated co-prescription designs, evaluated medication proceed by making association procedure learning, complex network examination, and cluster investigation [35]. In an existing study, the classification constructed DM methods are useful for healthcare data. It focuses on the estimation of heart disease by applying three classification methods that are KNN, DT, and NB [36].

An RF and SVM model were included in the current study's framework, which was augmented by the slime mold algorithm (SMA). SMA was enlisted to execute an optimal SVM model after RF identified the primary concerns. It is based on data from the coronavirus identity 2019 (COVID-19), including relevant studies involving RF-SMA-SVM and other well-known ML techniques. The results reveal that RF-SMA-SVM improves classification presentation and advanced consistency on four metrics, as well as the primary aspects screens that distinguish severe COVID-19 people from non-severe COVID-19 people [37]. The existing study assesses diabetes by the main features and classifies the relationships between inconsistent elements. The RF purpose delivered important feature boundaries. RF classifier explored the diabetes assessment. It was suggestions 75,7813 greater accuracies than SVM and may contribute to medical experts in making valuable decisions [38].

Another study primarily identifies the diabetes possibility DM method founded on the digital record of healthcare analysis, aims to deliver specific ideas and guidelines, which are conducted in the research trial. The investigational outcomes demonstrate that the normal forecast correctness of

the DT is 1.21%, and the outcomes of the training and test are the same, demonstrating the appropriate outcomes of the training set [39].

The primary goal of the DM system for healthcare is to give competent, thorough data point dissemination in datasets employing chaotic biogeography-based optimization and information entropy (CBO-IE) and also to prepare residents using chaos theory. Both evidence entropy & chaos theory permit the development of biogeography-based (BO) merging speed in the worldwide examination zone for more accurately picking group numbering and cluster participation. The CBO-IE was run on eight medical IoT datasets in a matrix laboratory, and the results show that the f-measure provides improved presentation [40].

In terms of accuracy, comprehension, positive possibility ratio, negative analytical value, and positive analytical value, the DT model outperformed DM approaches, according to the results of another research. As a result, the procedure is a good classifier for looking into factors that cause a delay between a burn and the start of burn patient care [41].

The existing study was started on the evidence of trials, to conduct a theoretical framework that was trained on how data will be managed to better facilities within the South African healthcare accommodations. The explanatory method was engaged and qualitative data was collected from obtainable works. Structuration theory was used to monitor the investigation of data. a framework was established from the results, primarily to monitor and improve data are kept, recovered, accomplished, and used for medical developed services [42][43].

Artificial intelligence (AI) was used with NB and RF classification process to categorize many infection datasets like diabetes, heart sickness, and cancer. Presentation examination of the infection data for individual algorithms is considered and matched. The consequences of experiments show, the efficiency of classification methods on a dataset and its complexity [44].

Because the bags begun are comparable, the creation of vital pharmaceuticals and healthcare equipment in multiple institutions tends to be frequent and parallel over a relatively long period. Regulation in the outline of requirements between substance sets is developed by computing the tendency of desired patterns and requirements using an algorithm (apriori association) of the dataset. 33.3 percent support and 85 percent confidence are the linked criteria. where the chemicals that appear

are substances with occurrence and relationships into consideration to certify the availability of medications and medical devices [45][46][47][48].

2.2.2 Image-based research

Image-based research method includes MRI, fMRI, x-rays, and CT scans. It contains health devices generating data. The variation of health imaging informatics discusses the clinical operation and provides further guidelines for scheduled clinical repetition. It is a detailed advance in health accomplishment machinery for dissimilar modalities. Importance in the requirement for well-organized health facts organization policies viewpoint of AI in vast medical data analytics [49][50]. The latest learning-based system is automatically making CT images from predictable TI-weighted MRI created by an RF regression through pitch-based automatic symbols to efficiently detention the link between CT examination and MR images. Reconstructed positron images of emission tomography (PET) consuming the PCT display fault well lower recognized assessment consistency of PET/CT representing in height numerical similarity [51].

Despite the variation of imaging trials with scanners, the method performed effectively, suggesting that it may be carried out by inexperienced individuals and is likely connected to undetected patient data. CNN may be able to track the acceptance of crucial MRI in a predictable implementation to aid inpatient evaluation and management [52]. A semi-supervised learning system was used to convert a novel MRI biomarker of moderate cognitive impairment (MCI) to Alzheimer's disease (AD), then combine it with age and sensitivity metrics around the themes, resulting in a collective biomarker [53]. The separation of images into the open beam, bone, and soft tissue sections is required for X-ray image improvement and many other medical image handling applications. The ML application resolves this problem, where consequences in robust and arranged inference [54][55].

Chest X-rays play an important part in the diagnosis of illnesses such as pneumonia. The imaging approach similarly distinguishes COVID-19. ML-based categorization of the obtained rich characteristic by ResNet152 with pneumonia & COVID-19 afflicted individual on chest X-ray is proposed in this present work. COVID-19 infection is further investigated in asymptomatic individuals using a non-aggressive and primary estimate of COVID-19 by investigative chest X-rays. [56][57]. To detect COVID-19 patients, a programmed revealing system called EMC Net was developed. A CNN was discovered to gain acceptability for the system to mine COVID-19 patients'

deep and high-level characteristics from X-ray images. For the identification of COVID-19, the ML classifiers (DT, SVM, RF, and Ada Boost) was identified [58].

DNNs are nowadays ML models in a range of areas, from image examination to natural language handling, widely arranged in the academic world and manufacturing. This progress has a massive prospective for healthcare imaging technology, health data analysis, health diagnostics, and general healthcare, gradually being understood. The author delivers a short outline of recent progress, and some allied challenges in ML realistic to health image handling and image examination. It is a very comprehensive and fast-increasing field, that applies specific emphasis on deep learning (DL) in MRI [59][60].

AI endures garnering considerable attention in medical imaging. The possible presentations are huge and comprise the sum of the health imaging life cycle after image construction diagnosis to result in prediction. The main problems to growth and clinical application are AI procedures including the obtainability of adequately large, curated, and descriptive training data that contains expert labeling. Existing supervised AI systems need a curtain procedure for data to train, validate, and test algorithms [61].

The CNN algorithm has a higher image resolution, a larger, more accurate lesion area detection and division, and is associated with older approaches. Post-operative imaging structures infection of bone and joint revealed three appearances: soft tissue blister, periosteum expansion, and bone damage. In the end, the research of imaging characteristics using X-ray, CT inspection, and MRI inquiry was enhanced. The sensitivity and specificity of X-ray were lower than CT/MRI. The testing results show that the CNN algorithm may successfully detect, disseminate, and aid clinicians in more quickly identifying abnormal images in hospitals [62][63].

The ML-generated classifiers can consistently discriminate COVID-19 patients' CXR images from other pneumonia techniques. This current work uses a dimensionality reduction strategy to construct a set of best categories for CXR images so that a well-organized ML classifier can differentiate COVID-19 possessions from non-COVID-19 possessions with surprising accuracy and sensitivity may be created. The ideal CXR images from spending time all around the world [64][65].

A new ML technique was suggested for sorting COVID-19 patients and non-COVID-19 patients on chest x-ray images into two groups. New negligible multi-channel booster moments are used to extract the categories from the chest x-ray images [66][67]. COVID-19, or COVID-CT-MD, is a new COVID-19 dataset that is also strong and covers community-acquired pneumonia (CAP). COVID-CT-MD dataset has improved lobe-level, slice-level, plus patient-level labeling, making the COVID-19 inquiry easier, and it can support the creation of progressive DNN and ML decisions [68][69].

2.2.3 Signal-based research

The signal-based research methods contain ECG and EEG. It mostly displays results in the shape of graphs and collects data from sensors. Suitable health devices produce large sizes of data that helped discover health risks. The procedure network patient's ECGs and put them on ML classifiers to categorize cardiac health dangers and assess strictness [70]. ECG is usually used for arrhythmia disclosure. The ML methods with signal treating algorithms usage for automated identification of heart health ECG used [71][72]. The current work aimed to develop a perfect ideal for identifying sleep stages using several types of HRV extracted from an ECG. The arrangement of the sleep phase can be used to estimate the proportion of sleep phases. Evidence of the proportion of sleep phases can provide insight into the utility of human sleep. To choose features and define the number of unknown nodes, a mix of exciting learning machines (ELM) with particle swarm optimization (PSO) has been used. The results were linked to SVM and ELM systems, which were less effective than ELM and PSO together. The accuracy ratings for the ELM and PSO as a whole were 62.66 percent, 71.52 percent, 76.77 percent, and 82.1 percent, respectively. The organization correctness can be better by organizing a PSO algorithm for feature selection [73].

The current work demonstrates a method for recognizing illnesses based on ECG and EEG data, using mobile sensors. The information was gathered through the elderly's habits, completion of the timed-up, go test, and various infections discovered during the study's trial [74][75]. The electrocardiogram (ECG) provides important information regarding some heart problems. The inquiry community's primary goal has been to understand and check for life-threatening cardiac problems. Signal handling methods, such as ML and its portions, such as deep learning, are widespread ways of evaluating and classifying the ECG signal, which is widely used for early detection and organization of cardiac conditions and arrhythmias [76].

Semi-supervised ML approaches are used to address the created signals since they provide a patient-specific strategy owing to key characteristics such as adaptability and forcefulness. There is a pressure metric proposed there that will provide mechanisms to halt and avoid potential chronic medical concerns for those who are extra sensitive, to each healthcare history [77]. Individual health continuous observation by wearable biomedical sensors is pleasant a type these days kits appropriate and easily accessible [78]. Scientists have happening to practice EEG for emotional acknowledgment. An emotion detection system based on period field arithmetical features [79][80]. The pyramidal one-dimensional CNN (P-1D-CNN) is a model system that has been studied previously. Even though a CNN system receives the internal architecture of data and outreaches hand-engineered procedures, the key difficulty is the enormous number of learnable difficulties. P-1D-CNN method on the knowledge of modification approach to reduce this issue, and it covers 61 percent less limitations compared to regular CNN system, and it has stronger generalization [81][82].

The method employed in a previous study to develop CADFES: computerized automated detection of focal epileptic seizures. K-NN, RF, adaptive boosting (AdaBoost), and SVM classifiers was used to evaluate the procedure's demonstration. For an automatic combination of focal & non-focal seizures, a software application known as CADFES was announced [83]. A construction for drowsiness discovery by physiological displays that present four assistances: First is crumbling EEG signals into wavelet sub-bands to mine additional marked facts far off raw signals, Second is mining and synthesis of nonlinear constructions from EEG sub-bands, third is a synthesis of the indication from EEGs and eyelid activities, fourth is using effective, particularly learning appliance for the status organization. The trial results show a great revealing accurateness but also a very rapid calculation speed [84][85]. Medical skilled components are an example of appropriate and intelligent medical detecting gadgets that are employed in everyday life. The combination of arrhythmic beats is commonly utilized to detect ECG defects to detect cardiac problems [86][87].

The studies presented in the current research show that using specific types of EEG biometric information exclusively while walking may accurately regulate stroke precursors as well as an occurrence in the advancing age by greater than 90% precision. Furthermore, an RF algorithm using quartiles and z-score regulation confirms the organization's scientific meaning plus presentation with a stroke estimate accuracy of 92.51 percent. The suggested technique in the existing study may be used at a low rate, and it can be beneficial for early infection diagnosis and estimation utilizing real-time stroke originator indications [88]. ML is becoming more well-known as a useful technique in

medical applications in epilepsy research. One of the most important applications of ML is seizure detection and estimation, which may be done using wearable sensors (WDs). The author discusses the use of WDs with ML in epilepsy, as well as prospective developments in these disciplines. There is evidence that epileptic seizures may be reliably detected using implanted EEG electrodes, non-EEG, and wearable sensors [89].

2.3 Machine learning algorithms

All three types of approaches outlined above employ ML. Many ML methods are utilized in the literature by various authors, which are detailed below: -

2.3.1 Linear regression

Linear regression is a process model which finds a link between a dependent variable and one or many independent variables by using a process-oriented approach. When there are simply constant independent variables, it is recommended [5]. One of the most used supervised ML methods is linear regression. The purpose of regression is to forecast a constant value, or in software words, a hanging number. A regression job is predicting the weather using weather forecasts or associated data from prior years [22].

2.3.2 Logistic regression

Logistic regression is a possibility-based technique wherein the cost function is a sigmoid function with a value between 0 and 1. There are two forms of logistic regression. Binary logistic regression is useful when the observations must be categorized into two classes, whereas multinomial logistic regression is used when the observations must be categorized into more than two classes. When a regression issue contains a dichotomous dependent variable, logistic regression is extremely useful [5].

2.3.3 Support vector machine (SVM)

A support vector machine (SVM) is a supervised ML model that uses classification algorithms for two-group classification problems. The notion of categorizing hyper-plane is used in this approach. The goal is to find a plane that splits the dataset is divided into two groups and maximizes the separation between both the sample points in each group. The range of this hyper-plane is reported to be the greatest. Data points that lie on opposite ends of such a hyper-plane are divided into groups. The hyper plane's size is related to the number of characteristics. The hyper-plane is just a line if indeed the number of features is below or equal to two. For three features, it transforms into a two-dimensional plane, but imagining it with more than three characteristics becomes challenging. SVMs have the benefit of being particularly resistant to over fitting concerns. SVMs will not only describe information by linear functions, and so they can also categorize datasets utilizing non-linear grains [1][5].

After giving an SVM model sets of labelled training data for each category, they're able to categorize new text. Compared to newer algorithms like neural networks, they have two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to have access to a dataset of at most a couple of thousands of tagged samples. The basics of SVM and it works with example. The two tags: red and blue, and dataset has two features: x and y. A classifier required, given a pair of (x, y) coordinates, outputs if it's either red or blue. Labelled training data on a plane will be draw. A SVM takes that data points and outputs the hyper-plane (which in two dimensions simply a line) that best separates the tags. That line is the decision boundary: anything that falls to one side of it that will classify as blue, and anything that falls to the other as red [6][22].

2.3.4 K-nearest neighbors (KNN)

KNN is indeed a supervised learning approach that classifies items based on their resemblance to specific attributes of other objects in a present category [1]. It is one of the simplest ML algorithms based on supervised learning technique. If similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. This algorithm stores all the available data and classifies a new data point based on the similarity. It means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

It can be used for regression as well as for classification but mostly it is used for the classification problems [5].

K-NN is a non-parametric algorithm**,** which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data [6]. KNN is a really simple ML technique. The method selects the position in the training examples that are nearest to the new data point to forecast a new data point. The labelling or outcome of this training example is then assigned to the new observation. The k in KNN denotes that any fixed number k of neighbors in the training set can be considered instead of only the nearest neighbor to the new data point (i.e, the nearest three or five instances). After then, the consensus forecast is allocated toward the new observation. Using just one neighbor for categorization may result in extremely high training sample correctness, but the system will be very complicated, with a large risk of errors. The number of neighbors used for categorization must still be specifically selected such that perhaps the k-NN system is hardly too complicated or even too easy to generalize. Developing the k-NN system is normally quick, but a prediction might take a long time if you have a big training dataset (whether in terms of content or observations) [22].

2.3.5 K-means

Objects are classified using K-means depending on whether or not there fit within the constraints of a specific class. As a result, the categorization possibilities are confined to "similar" versus "dissimilar". The center of the group for every class is determined using Euclidean distances, as well as a unique update is simply categorized based on its range from every group. This technique is used in a variety of web browsers & wireless sensor network (WSN) technologies. Mobile sensing networks are utilized to identify injuries amongst soldiers distant from respective positions during a conflict, and data acquired by portable IoT nodes are often used to analyze the ECG of victims. [5].

2.3.6 Decision tree (DT)

Because the model is simple to grasp, the DT is the most often utilized DM approach. A DT is a network with a root node, branch, and leaf nodes. The DM methods that have been used to diagnose

and prognostic breast cancer on such a validation set as well as the SEER set of data, the DT was determined to become the strongest indicator, with a 93.62 percent efficiency [6]. A database DT is a tree-like representation of data. It is used in industrial engineering and ML to make decisions that lead to valid conclusions, as well as in DM to retrieve insight from clinical evidence. It is comprised of three sorts of nodes: choice nodes, chance nodes, and end nodes [1][5].

2.3.7 Random forest (RF)

RF is a community model in which several trees are grown and objects are a classified number of votes of all of the other trees [6]. The RF method is a model of several decision trees, each being dissimilar from others, as the term 'forest' inside the term indicates. The RF prototype is based on the idea that every tree may do a great job of estimation, and will over fit on a portion of such data. By aggregating the outcomes of several trees that all cooperate well and fit the data in various ways, the number of errors can be limited. Many decision trees must be developed to apply the approach behind the notion of random forests. Every tree must be distinct from the others and perform an excellent or good job of ensuring the objective [22].

2.3.8 Naive bayes (NB)

The Naive Bayes (NB) classifiers come from a family of models which are very close to linear models. However, NB classification methods are quicker in training. To accomplish this efficiency, NB models usually have somewhat inferior generalization capability than linear detectors like logistic regression. This is performed because NB trains constraints by examining every component separately and gathering basic per-class data from every characteristic [22]. The Bayes method is used to conceptualize NB categorization. The term 'naive' assumes that almost all characteristics are distinct from one another. A feature matrix or a response vector are created from the data. The feature matrix's rows reflect that this whole data set is in form of the feature vector, which together means a unique parameter. Each row of the answer vector, but on the other side, indicates a result category [5][6].

2.3.9 Adaptive boosting (Ada-boost)

There in the case of weak trainees, in particular, the performance is comparable to that of a randomized result producer. As a result, combining them through many ML algorithms to generate a strong learner is a suitable method to use them. The cluster method is learning for training model that employs several learners. Appears to be a suitable learning strategy that distributes weights to every weak learner based on how the borders are identified or predicted in the data. This process is done until the model is adequate. Ada boost starts by assigning equal weight to each view (for the initial boundary), then gradually increases the values for improperly categorized items and adjusts the borders appropriate unless all insights are properly categorized [5].

2.3.10 Convolutional neural networks (CNN)

CNN is just a feed-forward network that is used to classify data. It decomposes the information into components and then sends them to a convolution layer, which combines those parts in various ways until response to the growing (convolution). The input images are then mapped against such sequences by a rectified linear unit (ReLU) layer, which finally transfers them onto another convolution layer [5]. The CNN was created to hunt for a system that can simulate a brain. The CNN was created to solve ML classification issues using a non-linear premise. Among the most significant benefits of CNN is its ability to extract information from massive volumes of data and develop extremely complicated models. If given sufficient time, data, and good parameter adjustment, CNN may typically outperform other ML methods [22].

2.3.11 An Artificial neural network (ANN)

An ANN is an ML method that simulates the human brain's educational process, including input data which takes the data for processing, many layers that analyze the data, and output units that show the results. The hidden units in ANNs accept input materials, apply a randomized value and bias each, and compute numerous weighted amounts, that was then transferred via stages with values and totals until eventually achieve the ultimate layer, which determines the output using a non-linear- linear activation. Whenever the results are wrong, a functional form feeds them back to the earlier layers (back-propagation) to change the values until accurate responses are produced. ANNs

are exceedingly adaptable and gain use in pattern matching areas [5]. An ANN is a biological neural system with the capacity to learn and interpret data streams. The activation function, which could be linear or non-linear, determines the network connectivity and efficiency. The usage of ANN can aid in predicting the optimum approach to patient care symptoms. [1].

2.3.12 Bayesian classifier

In the realm of medicine, the Naive Bayes classifier is a useful tool. In health research, a Bayesian Classifier is a conceptual procedure for making a health diagnosis that is predicated on the probabilistic theory and may be utilized in computerized health diagnosis expert systems. It can deal with an unlimited number of parameters, both categorical and continuous [1].

2.4 Summary

Many individuals have concentrated their energy in the DM sector, notably in health treatment, for the ease and improvement to donate to the health industry for such benefit of human health, according to current research. In table 2.1, a few studies are mentioned below:-

Table 2.1: Existing studies summary of the literature review

Reference	Title	Method	Contribution	Advantages	Limitations
[49] A.S. Panayides 2020	Current Issues and Future Trends in AI and Radiology Technology.	Remote Healthcare Applications: ECG Output Processing and SVM Classifier-Based Abnormality Detection Disease categorization and organ/tissue segmentation, with an emphasis on AI & ML architectures, have now become a de-facto methods.	Integrative analytics techniques powered by affiliated research branches described in an existing study have the potential to change imaging informatics as we know it today both for radiology & digital pathology applications across the healthcare continuum.	The therapeutic advantages of developments in in silico modeling connected to emer ging 3D reconstruction and v isualization applications are shown further.	Imaging researchers have to deal with handling large, multi-dimensional sets of data and properly interrogating the features of data from many senses.
[51] Xiaofeng Yang 2018	Histologic signal and ML-based reduction adjustment for brain PET/MRI.	For contrast enhancement of brain PET, a learning-based technique to build client CT maps from regular T1-weighted MRI in its native space.	This paper describes a learning-based technique for generating patient-specific CT maps in native space from standard T1-weighted MRI for brain PET attenuation correction.	Reconstructed PET images for a brain scan utilizing the PCT have errors that are considerably below the approved PET/CT test reliability, demonstrating strong quantitative equivalency.	To assess the viability of this strategy, a small number of patient datasets was used in a previous study.
[52] Silvia Basaia 2018	A combined MRI was used to classify Alzheimer's disease and moderate cognitive impairment.	Individual diagnosis of Alzheimer's disease and moderate cognitive impairment is predicted by a deep learning system.	Focused on a personal cross-sectional brain functional MRI scan, the author developed and evaluated a deep learning method for identifying the individual treatment of AD and	The CNN show has the potential to create a paradigm for the automated, individual, and early diagnosis of Alzheimer's disease, hence speeding up the use of	It cannot rule out the possibility of future c-MCI in s-MCI patients.

			moderate cognitive impairment that would transition to AD (c-MCI).	structural MRI in normal practice to aid in patient evaluation and management.	
[53] Elaheh Moradi 2015	MRI-based Alzheimer's conversion diagnosis in MCI patients using a computational model.	A new MRI-based technique for predicting the progression of MCI to AD.	A new MRI-based approach for predicting the change of MCI to AD one to 3 years well before clinical diagnosis.	When MRI data was combined with aging and cognitive assessments, the AD transition prognosis in MCI patients was dramatically improved.	The use of unlabeled data from uMCI participants in the LDS learning technique somewhat improved classification performance, but not enough to attain statistical significance.
[54] Joseph Bullock 2019	XNet is a limited CNN solution for medical X-Ray imagery categorization.	A final ML solution to this issue produces robust and effective prediction.	The author takes an ML approach to this problem, giving a top strategy that produces efficient and reliable reasoning.	The solution achieves a 92 percent accuracy rate, an F1 value of 0.92, and also an AUC of 0.98, outperforming traditional image processing algorithms.	Given a modest sample, a completely automated approach for segmenting medical X-Ray scans.
[56] Rahul Kumar 2020	COVID-19 Detection utilizing SMOTE & ML Classification methods on Chest X-Ray Imaging.	On chest X-ray scans, current research provides an ML-based categorization of the obtained deep component utilizing ResNet152 using COVID-19 and Pneumonia cases.	Employing COVID-19 and Pneumonia cases on chest X-ray data, this previous study provides ML-based assessment of the derived deep characteristic using ResNet152.	The construction of this strategy would be beneficial in predicting the epidemic earlier on that will assist in its successful containment.	The data was collected in an early stage of COVID-19, which is a restriction.
[58] Prottoy Saha 2020	EMC Net: COVID-19 detection from X-ray imaging to use a CNN and an ensemble	By analyzing chest X-ray data, the EMCNet automatic recognition system was presented to detect COVID-19 individuals.	By analyzing chest X-ray images, the EMCNet automatic detection technique was presented to find COVID-19	With 98.91 percent accuracy, 100 percent precision, 97.82 percent recall, and a 98.89 percent F1-score, EMCNet	Although EMCNet has significant limitations (for example, it may misclassify some COVID-

	of ML classification models.		cases. To capture deep and high-level information from X-ray images of COVID-19-infected patients, a CNN was created to keep model simple.	outperformed other contemporary deep learning-based methods.	19-positive patients as negative), it may be used as a substitute for manual radiological analysis and can help clinicians discover COVID-19 from chest X-ray scans automatically.
[64] Abolfazl Zargari Khuzani 2020	COVID-Classifier is a ML method that may be used to help diagnose COVID-19 virus in chest x-ray data.	An effective ML classifier that could differentiate COVID-19 is built using an image compression approach to construct a collection of ideal features from CXR images.	An effective ML classifier can reliably identify COVID-19 CXR images from normal cases and pneumonia generated by other infections, according to a previous research.	It spreads over the X-ray image's full chest region. Image characteristics are computed both in spatial and frequency domains (Texture, GLDM, GLCM) (FFT and Wavelet).	The CXR sample is modest in comparison to other datasets.
[66] Mohamed Abd Elaziz 2020	Image-based COVID-19 assessment using a novel ML approach.	The computational procedure is sped up by using a multi-core parallel processing architecture.	To extract features from the COVID-19 x-ray images, the approach used a fractional moment (i.e., FrMEMs).	The method achieved accuracy rates of 96.09% and 98.09% for the first and second datasets, respectively.	Due to resource constraints, existing research compares the suggested paradigm to Mobile-Net. On both datasets, a comparative with Mobile-net & related studies.
[68] Parnian Afshar 2020	A model COVID-CT-MD: COVID-19 CT Scan Dataset (ML and DL).	COVID-CT-MD can lead to the development of sophisticated ML and DNN solutions.	This existing work introduces COVID-CT-MD, a novel COVID-19 CT scan collection	COVID-CT-MD can help with the creation of sophisticated ML and DNN systems.	Rather than minor results, the slice & lobe labeling procedures produce unique

			that includes more than only COVID-19 instances.		manifestations. Because of their poor differentiation, several worrisome spots next to the chest wall & diaphragm are not designated as "infected."
[70] Shurouq Hijazi and Alex Page 2016	In heart health tracking and decision-making, ML is being used.	To discover heart health issues and assess severity, the technique filters participants' ECG and uses ML algorithms.	To discover heart health issues and assess severity, the technique filters patient ECG and using ML models.	The employment of a voting predictor, which seeks to combine the estimates of numerous different classifiers in order to get a superior outcome, is one option.	The processes might be used to a broad variety of medical data & disorders. The improvement of our technique and the expansion of EHR databases are expected to considerably enhance the quality of patient care with a range of diseases.
[71] Turker TUNCER 2019	Using a unique hexadecimal local pattern and multilayer wavelet transform with ECG data, an automated arrhythmia detection system was developed.	Arrhythmia identification utilizing ECG readings using traditional ensemble mining & DL approaches.	To use a 1NN predictor and city block suitable metrics, the current system was able to define 17 arrhythmia classifications with 95.0 percent accuracy.	For arrhythmia identification utilizing ECG data, the present technique outperforms competing methods such as traditional ensemble learning & DL.	The method's shortcoming is that each of the 17 classes uses a smaller dataset.
[76] Muhammad	A Survey of Stages-Based ECG Signal	A CNN but a generalised minimax-concave (GMC)	The authors used MITDB to test their one-dimensional ECG data	One or even more levels of the proposed model could be	In the reviewed studies, a 2-D image-based

Wasimuddin 2020	Signal Analysis Through Traditional Signal Processing to DM Algorithms.	algorithm are used for the algorithm.	processing and classification algorithms.	classified as a phase paradigm.	categorization of ECG receives very little consideration.
[78] Mr. Amit Walinjkar 2017	Authentic ECG analysis & medical integration with personalized sensor devices.	T Utilizing k-NN classifications, ML models are trained upon that MITDB arrhythmia database (MIT-BIH Physionet) achieved accuracy of above 97 percent.	The learning techniques that was trained with the MITDB arrhythmia dataset produced 97 percent and greater accuracy in prediction and classification.	In the categorization and prediction of four forms of arrhythmia, supervised learning models based on the six characteristics outperformed unsupervised learning techniques (V, A, L, R annotations in MITDB).	An installation of a HAPI-FHIR production servers showed the real-time recording of ECG readings to EHR for subsequent analysis by medical doctors and medics, using a standardized SNOMED coding scheme.
[79] Fabian Parsia George 2019	EEG data are period examined as well as an SVM method is used to recognize human emotions.	A method for detecting emotions based on statistical data in the time-frequency domain. The best features are chosen using a box-&-whisker plot, these are then input into an SVM classifier to train and evaluate the DEAP dataset.	The DEAP dataset's heavily processed EEG waves was used to categorize two categories of emotions: valence & arousal.	The suggested strategy has a 92.36 percent accuracy for the evaluated dataset, according to the experimental findings.	When compared to earlier techniques for such a DEAP dataset, this model produces superior results. It is an emotional assessment database.
[81] Ihsan Ullah 2018	Rooted ML method, an automatic method of detecting epilepsy employing EEG brain waves.	A collection of P-1D-CNN models.	The system is made up of a collection of experience & basic (P-1D- CNN) algorithms that are fed a ECG Signal.	The P-1D- CNN approach may be used to construct strong intelligent machines for a variety of	Because the P-1D-CNN uses 61% less storage and store, its capacity and memory needs for wearable

				illnesses, not just epilepsy detection.	technology may be an issue.
[83] Shivarudhrappa Raghu 2018	Using neighborhood method and ML methods, we was able to distinguish between focal and non-focal EEG data.	CADFES, a new tool, was introduced.	Using an SVM with a cubic kernel function, simulation results show the maximum sensitivity, specificity, accuracy, positive predictive rate, negative predictive rate, and AUC of 97.6 percent, 94.4 percent, 96.1 percent, 92.9 percent, 98.8 percent, and 0.96 percent, respectively.	An approach aids neurologists in healthcare decisions by minimizing manual categorization mistakes and allowing them to treat additional patients.	The dataset 41.66 was utilized, as well as the ML method.
[84] Lan-lan Chen 2015	Wavelet-based nonlinear characteristics and ML are used to determine alertness/drowsiness through physiologic inputs.	Drowsiness detection system based on physiological cues.	The experimental findings reveal that the suggested technique not just to obtains a good performance as well as a quick computing time.	Decomposition of EEG data into wavelet thread to obtain more obvious information over raw signals, extraction & fusing of nonlinear characteristics from EEG sub-bands, integration of information from EEGs with eyelid movements, and status categorization using an efficient deep learning machine.	In certain ways, it may cause the body to bulky and impede activity. Second, this study only used one task to induce spontaneous drowsiness: persistent mental computation.

[86] C.VENKAT ESAN 2018	Anomaly Identification in Distant Smart Healthcare Using ECG Signal Editing & SVM Classifiers.	To differentiate between normal and abnormal participants, ECG data preprocessing with SVM-based arrhythmic pulse categorization are used.	To differentiate people into normal and abnormal, this work uses ECG signal preprocessing and SVM-based arrhythmic beat classification.	The SVM-based classifier's testing results show that it can distinguish between normal and arrhythmic hazard disordered people with a high precision of 96 percent.	Only a few available methods are time-consuming and need complicated calculations. When dealing with noisy data, morphological ECG characteristics are not possible.

CHAPTER 3

METHODOLOGY

3.1 Overview

Kitchenham guidelines [17] are used in the proposed study to perform SLR, with a focus on the software engineering area. Researchers must describe current material on any issue in an objective and thorough manner. An SLR may lead to more general conclusions regarding a given occurrence than single studies could, or it could serve as a precursor to future study [17]. These three primary steps of SLR are discussed in this chapter: planning, conducting, and reporting the review. The research practice and research design are discussed, as well as the use of stages to improve research techniques [4][17]. To begin, the planning phase includes basic research keywords related to the topic, the development of research questions, search strings for finding relevant existing studies, electronic databases or digital libraries for downloading research papers, and the use of inclusion/exclusion criteria for appropriate research articles. The second phase is review conduction, which involves data extraction and quality assessment. The third step of results reporting entails extending the authoring of findings to include organising and calculating the SLR.

3.2 Research strategy

SLR is a method of investigating diverse research in a methodical manner. After obtaining valuable information, the first chapter specifies the study questions. There are quantitative and qualitative findings [4]. SLR's main outputs are a collection of all relevant previous studies focused on research concerns. The establishment of a framework and guidance for future directions is also aided by the research approach [17].

3.2.1 Quantitative research strategy

Quantitative research involves systematic data exploration, theoretical methodologies, and scientific computations. Its goal is to expand on a previously expressed topic (Deductive Reasoning).

It makes an effort to limit the researcher's influence on the outcome and involves a lot of hypothesis analysis. Quantitative data encompasses statistical data and data collecting, necessitating closed-end responses [4].

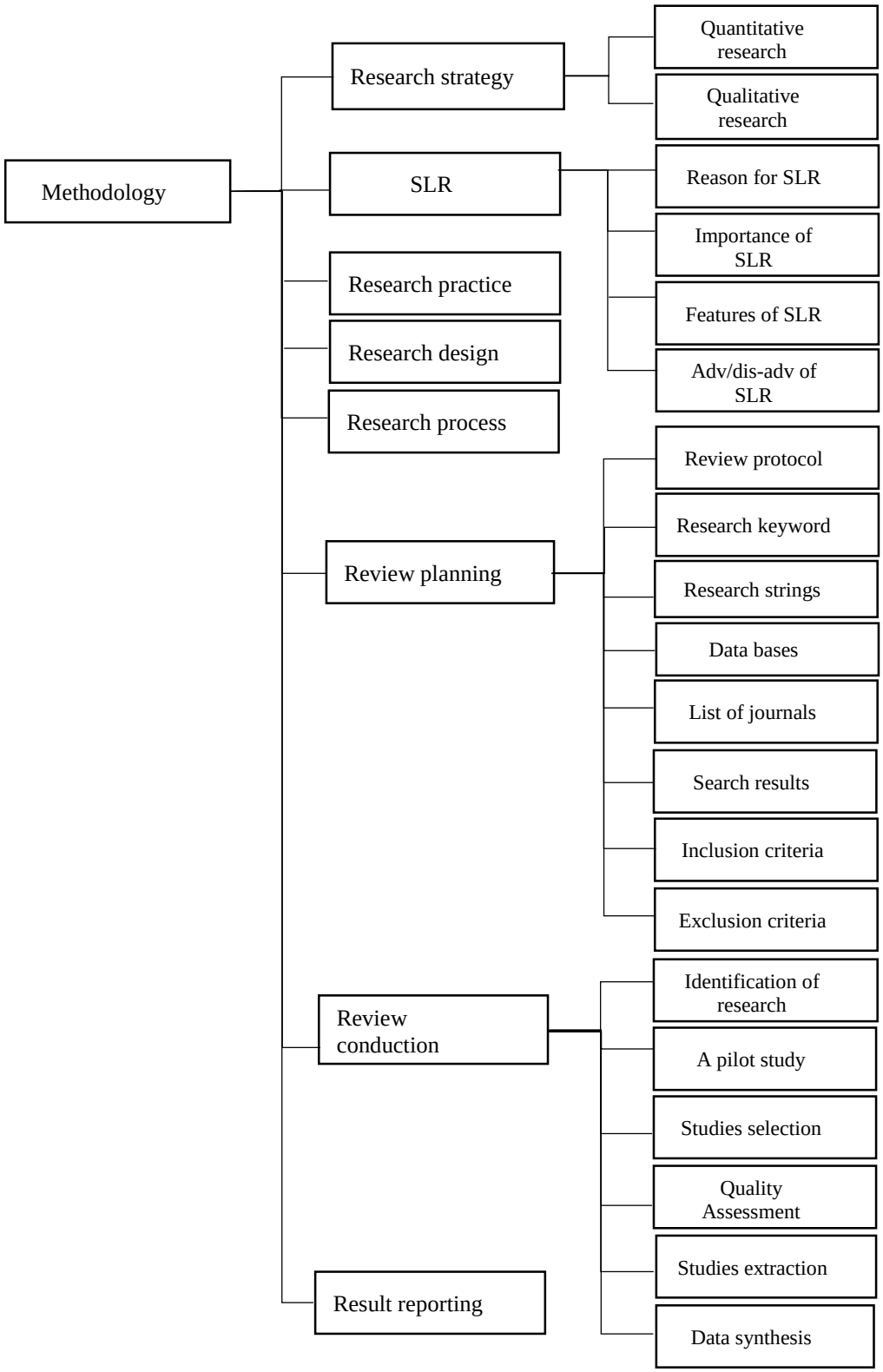

Figure 3.1: Layout of the research methodology chapter

3.2.2 Qualitative research strategy

Qualitative research yields non-numerical yet significant information for outcomes [17]. Its investigation attempts to generate a new concept (Inductive Reasoning). The researcher is an integral aspect of the study and plays an important function. Qualitative data imply data gathering and difficult reports or conclusions, therefore their authorizations are open-ended.[4].

3.3 SLR putting into practice

Assessing, identifying, and interpreting all existing research relevant to a certain research topic, breadth of thought, and subject area is what a systematic review, or SLR, is all about. Primary literature is a single study that underpins a systematic review, while secondary literature is a systematic review in and of itself [17].

3.3.1 Reasons for performing SLR

The SLR is performed due to many reasons and the most commonly used reasons are mentioned below:

a. It is required to summarize the current idea for implementations or knowledge e.g. and review the experimental confirmation for help and agile method applications with in margins.
b. The gaps are finding from recent research and to indicate area for important exploration
c. It bring framework or appropriate model for further novel research happenings.

3.3.2 The Importance of SLR

Every research has some kind of literature review or related work. SLR have scientific worth when it is perfect, detailed and reasonable. This is major motivation for choosing SLR. It is synthesizes current literature in a reasonable manner and emphasise it must be fair. For example, it is must begin with a predefined plan. The search plan restricted to measured completeness of search. In certain, researcher executing without support to their desired theory, they perform it with determination to identification and report investigation.

3.3.3 Advantages and disadvantages

The SLR have some advantages and disadvantages are mentioned below:

a. If the primary studies having publication bias then it cannot consider the results of literature so it creates well-defined methodology.
b. Empirical and comprehensive range methods provided information about any occurrences. The phenomenon is strong and convenient when studies provided trustworthy results. The causes of variance can be considered weather studies having inconsistent results.
c. The meta-analytic techniques combine data in quantitative studies. It is discovering genuine effects which are not detect by smaller studies.
d. The SLR required more struggle than other old-fashioned literature review methods. It is main disadvantage of SLR. The meta-analysis of enlarged influence also a disadvantage, In the meantime it is prospective to notice minor biases as well as correct assets.

3.3.4 Features of SLR

The SLR is segregated by a number of features form old method literature review, few of those are mentioned below [17]:

a. Defining a review protocol started the systematic review that identifies to address the research question and the approaches used for accomplishment of review.
b. Systematic review based on well-defined approach that goals to identify the appropriate works as conceivable.
c. Search strategy recorded by systematic reviews so that precision evaluated by readers, completeness as well as repeatability of process. Digital libraries are no duplicate searches.
d. Primary study is measured by inclusion/exclusion criteria in systematic review.
e. SLR require facts and figures achieved from every primary study with quality conditions.
f. Quantitative meta-analysis is required in a systematic review.

3.4 Research practice

In SLR planning, conducting, and reporting was accessible some strategies for a systematic review.

a. A protocol is developing.

b. The research question must be defined.
c. A researcher report the specific problem relating inclusion/exclusion conditions and responsibility of data extraction.
d. The search strategy is defining.
e. Quality include of each primary study and defining the data to be extracted.
f. Lists of studies which are included and excluded to be maintaining.
g. The procedures are used for data synthesis.
h. The guidelines are used for reporting of result.

3.5 Research design

For the purpose of finding connected studies, research questions and search strings are developed, which display the search query's associated results.

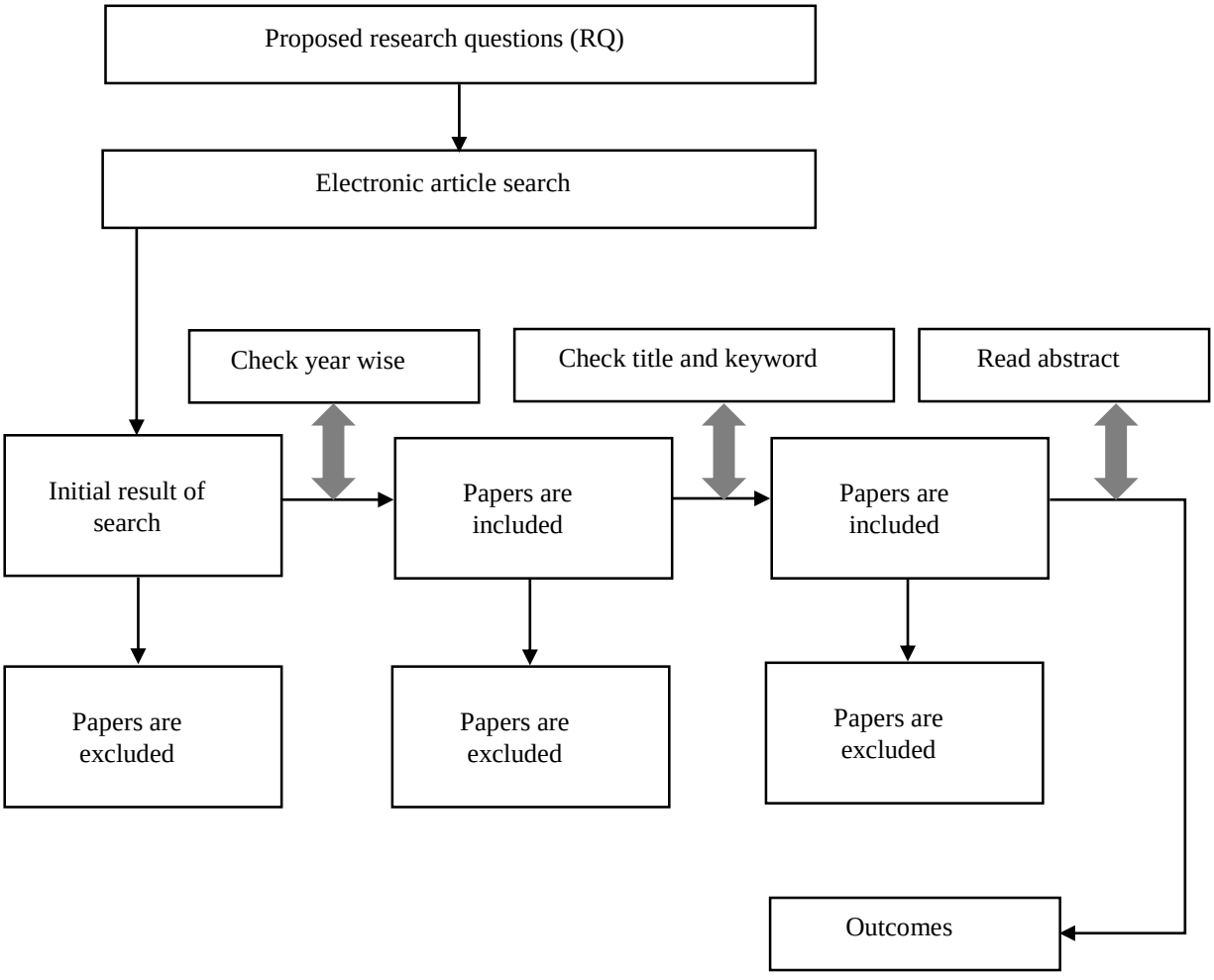

Figure 3.2: Detail flow chart of SLR

This study's search string aids in the discovery of DM, ML algorithms, and data gathering techniques that can help to enhance the research. Furthermore, the search query aids in the creation of the DM structure as well as its strength. The multidisciplinary digital publishing institute (MDPI) digital library is denoted as MD, Science direct as SD, the institute of electrical and electronics engineers (IEEE) as IX, Springer as Sp, and Other digital libraries like Hindawi, ResearchGate, Google Scholar etc as OL. Digital platforms are used for search and are denoted with identification in this document. Advanced search techniques such as the Boolean operator, brackets, and the AND operator are used to get more accurate results.

3.6 Research process

A number of discrete activities involved in a SLR. Activities arrangement and numbers are differently suggested by existing procedure of SLR. But the main stages of process are mostly agreed in the medical strategies. In the used methodology SLR have three major parts like Planning of review, conduction of review and report the outcomes found from this process. The process shows indication in below is figure 3.3

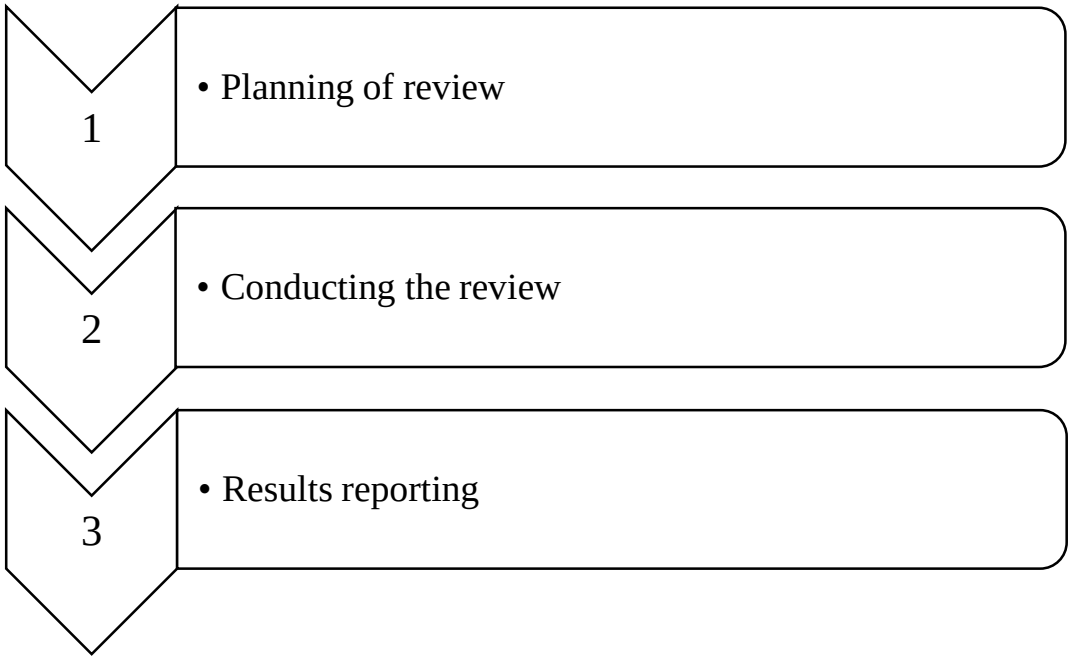

Figure 3.3: Three main phases of SLR

The phases related through planning of the review exist like requirement of a review are research questions identification, review protocol development and review protocol assessment. The

phases linked with the review conduction are in form of research identification, collection of initial studies, quality assessment of studies, data extraction from selected studies, monitoring of data and synthesis the data. The steps connected with the review reporting are mechanisms of dissemination, report formatting and the report evaluation. It is optional that the review protocol evaluation, the outcomes report evaluation and the systematic review team decided that by the quality assurance measures.

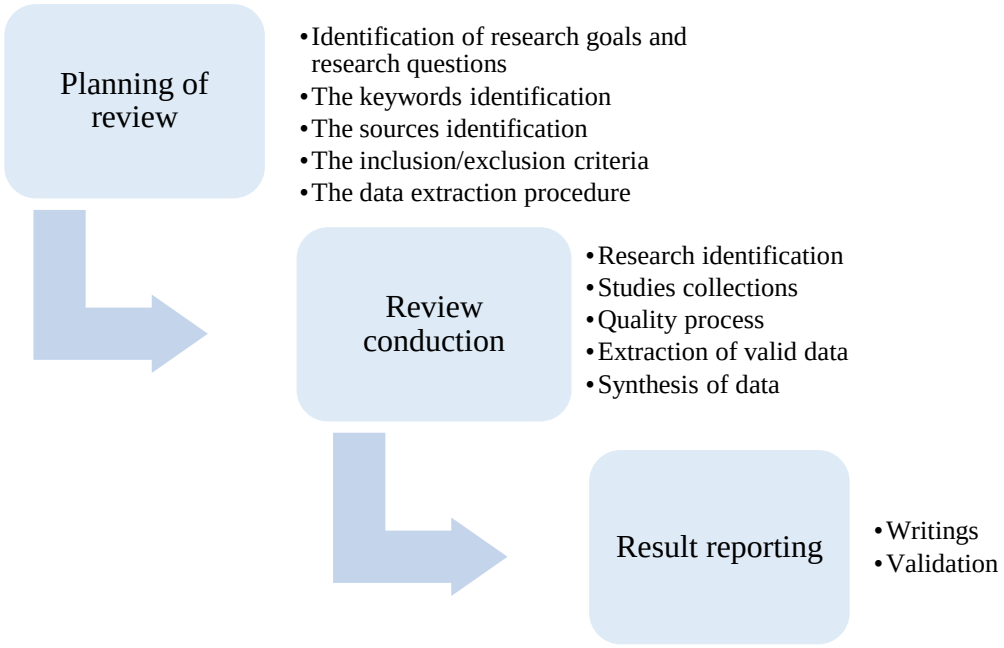

Figure 3.4: Complete steps of SLR phases

The phases definite above possibly will look to be in sequence, but iteration involves in many stages. In particular, the protocol improvement phase initializes many activities and when the review accurately accepted it may be refined. For instance, the inclusion and exclusion norms are made collection of initial studies. These norms are primarily stated when the protocol is enrolled but quality criteria is defined it may be polished. When the quality criteria was finalized, the protocol will be amended and forms of data extraction are initially prepared. It is possible that amendment took place in the data synthesis methods which is defined in the protocol [4][17].

3.7 Review planning

It is required to indorse the need for a review before undertaking a SLR. However, the defining research question are most significant pre-review activities that will report SLR and creating a review

protocol or plan describing the initial review events. The evaluation process of review protocol should also be substance to an independent [17].

3.7.1 Review protocol

A detailed systematic review undertaken by the methods of review protocol. The researcher bias was reduced by the pre-defined protocol. The parts of protocol included every single review features are listed below in figure 3.5

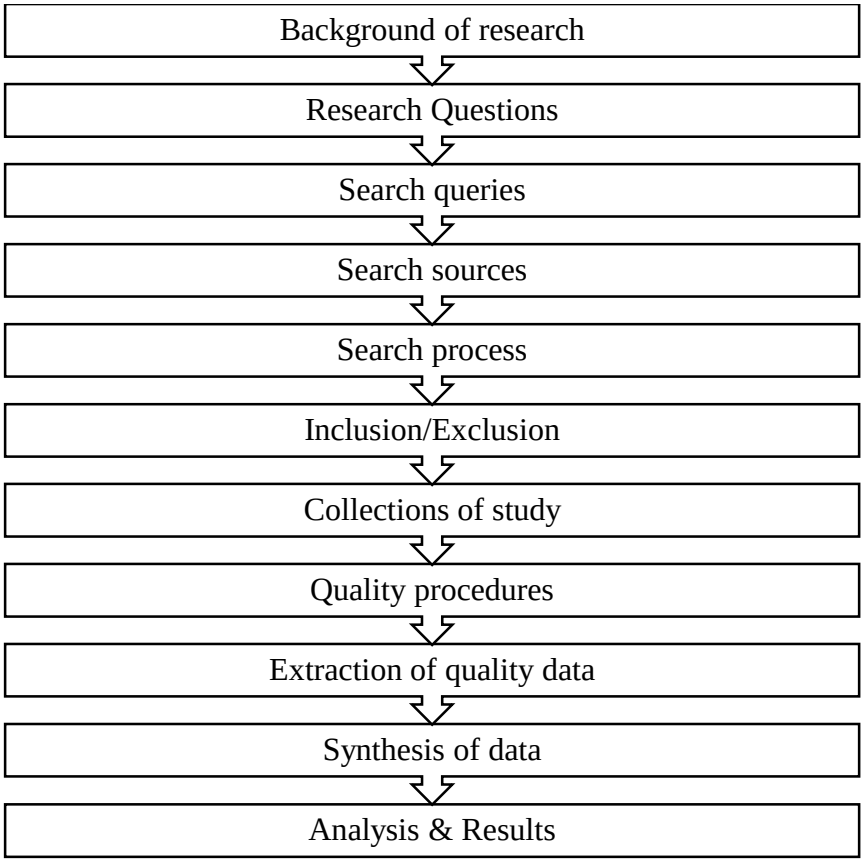

Figure 3.5: Protocols of SLR

In a systematic review, the protocol is a serious aspect. The protocol presenting to supervisors for review and criticism as for its evaluation is concerned.

3.7.2 Research Keywords

Research keywords are extracted from the title of the research and formulate the research questions which are pre-defined in the first chapter section 1.4 as well. These keywords are mentioned below.

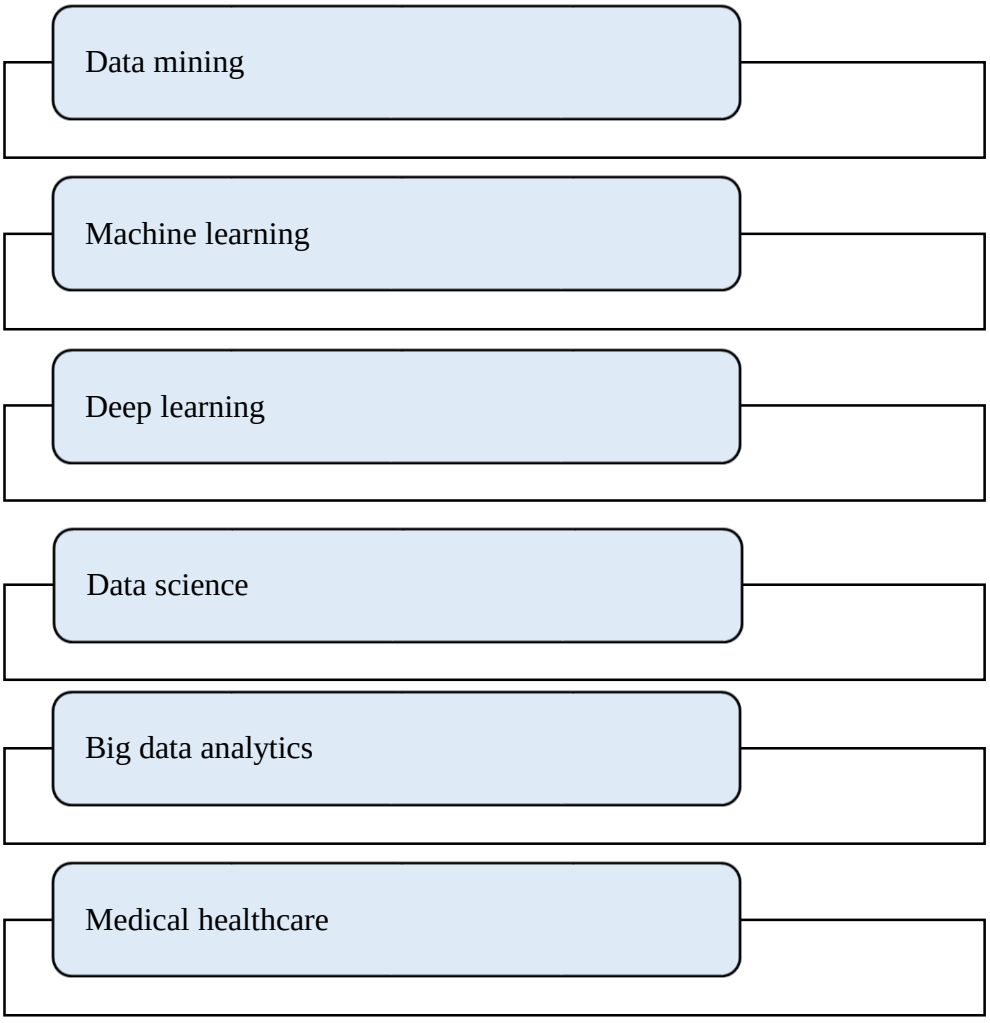

Figure 3.6: Keywords of proposed SLR

3.7.3 Search queries

The formulated research keywords developed further research queries to search related papers from electronics databases. Research strings or queries mentioned below in figure 3.7

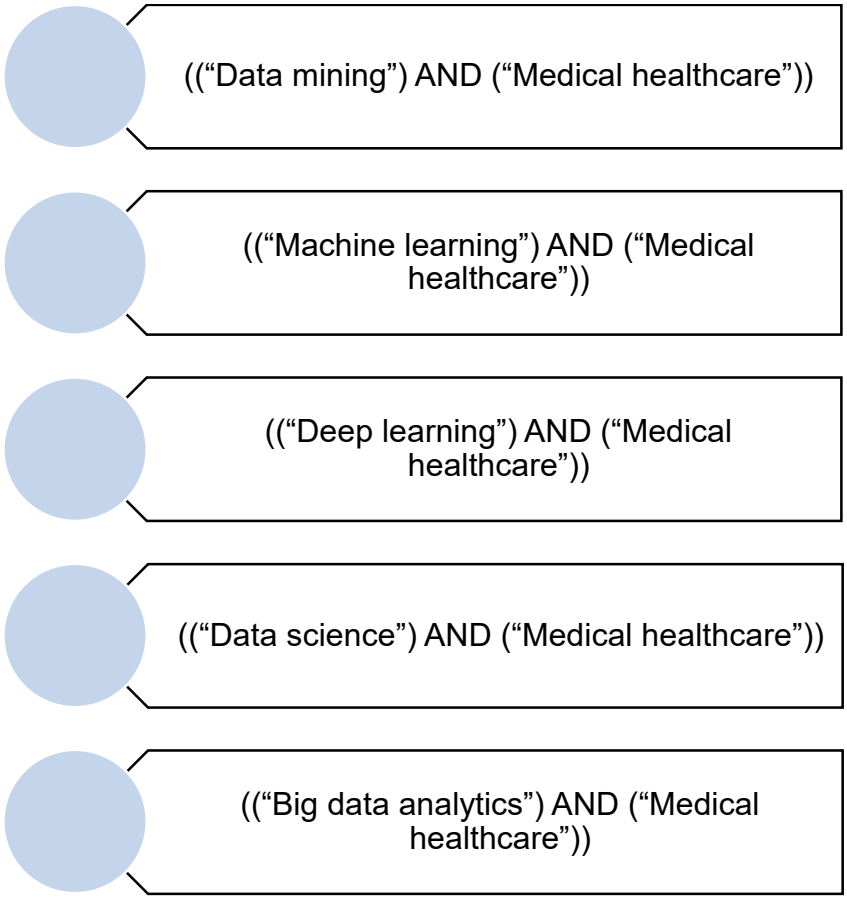

Figure 3.7: Research studies searching queries

3.7.4 List of databases

The undermentioned electronic sources achieve a comprehensive search which is relevant to software engineering and denoted identification in this document.

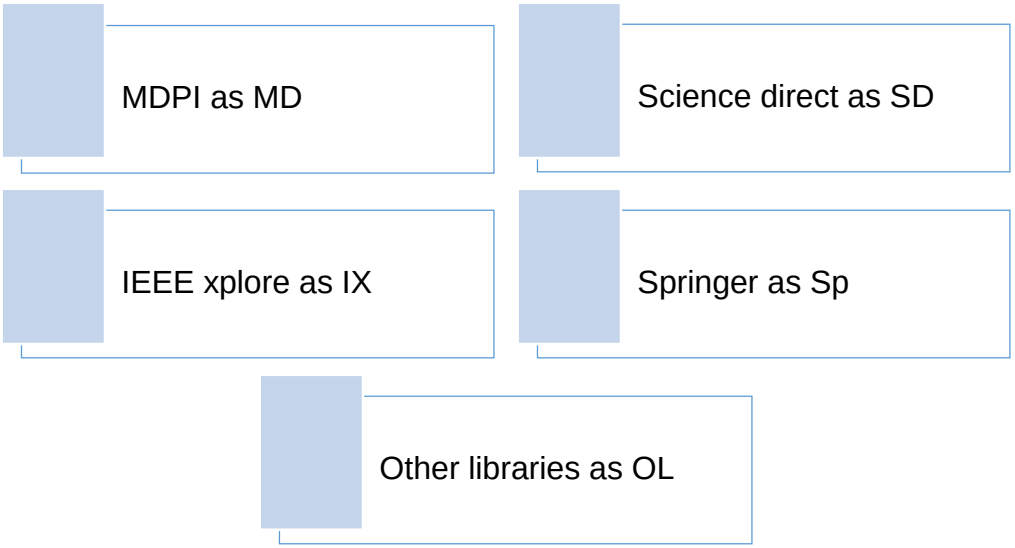

Figure 3.8: Electronic search databases

3.7.5 Search results

The search queries was used separately in above mentioned electronic databases. The entire results of the search queries are displayed in the tables below.

Topic: Role of data mining

Keyword 1: Data mining

Search query: ((“Data mining”) AND (“Medical healthcare”))

Table 3.1: Found search results of keyword 1

Databases denoted identity	No of results
MD	64
SD	72
IX	185
Sp	52
OL	1190
Total	1563

Topic: Machine learning

Keyword 2: Machine learning

Search query: ((“Machine learning”) AND (“Medical healthcare”))

Table 3.2: Found search results of keyword 2

Databases denoted identity	No of results
MD	151
SD	125
IX	460
Sp	93
OL	2050
Total	2879

Topic: Deep learning

Keyword 3: Deep learning

Search query: ((“Deep learning”) AND (“Medical healthcare”))

Table 3.3: Found search results of keyword 3

Databases denoted identity	No of results
MD	105
SD	74
IX	386
Sp	52
OL	1210
Total	1827

Topic: Data science

Keyword 4: Data science

Search query: ((“Data science”) AND (“Medical healthcare”))

Table 3.4: Found search results of keyword 4

Databases denoted identity	No of results
MD	29
SD	22
IX	1258
Sp	21
OL	459
Total	1789

Topic: Big data analytics

Keyword 5: Big data analytics

Search query: ((“Big data analytics”) AND (“Medical healthcare”))

Table 3.5: Found search results of keyword 5

Databases denoted identity	No of results
MD	50
SD	32
IX	72
Sp	19
OL	593
Total	799

3.7.6 Inclusion criteria

The DM in medical healthcare is the leading effort is to find out the studies. For this resin, the detailed inclusion and exclusion criteria are fixed to achievement superior consequences agreeing to the dedicated topic. The titles of research studies, keywords, abstract are find out appropriate research with the help of research questions and inclusion criteria. The existing studies are in below categories.

a. Language must be English.
b. Discuss DM in medical healthcare.
c. Discuss ML in medical healthcare.

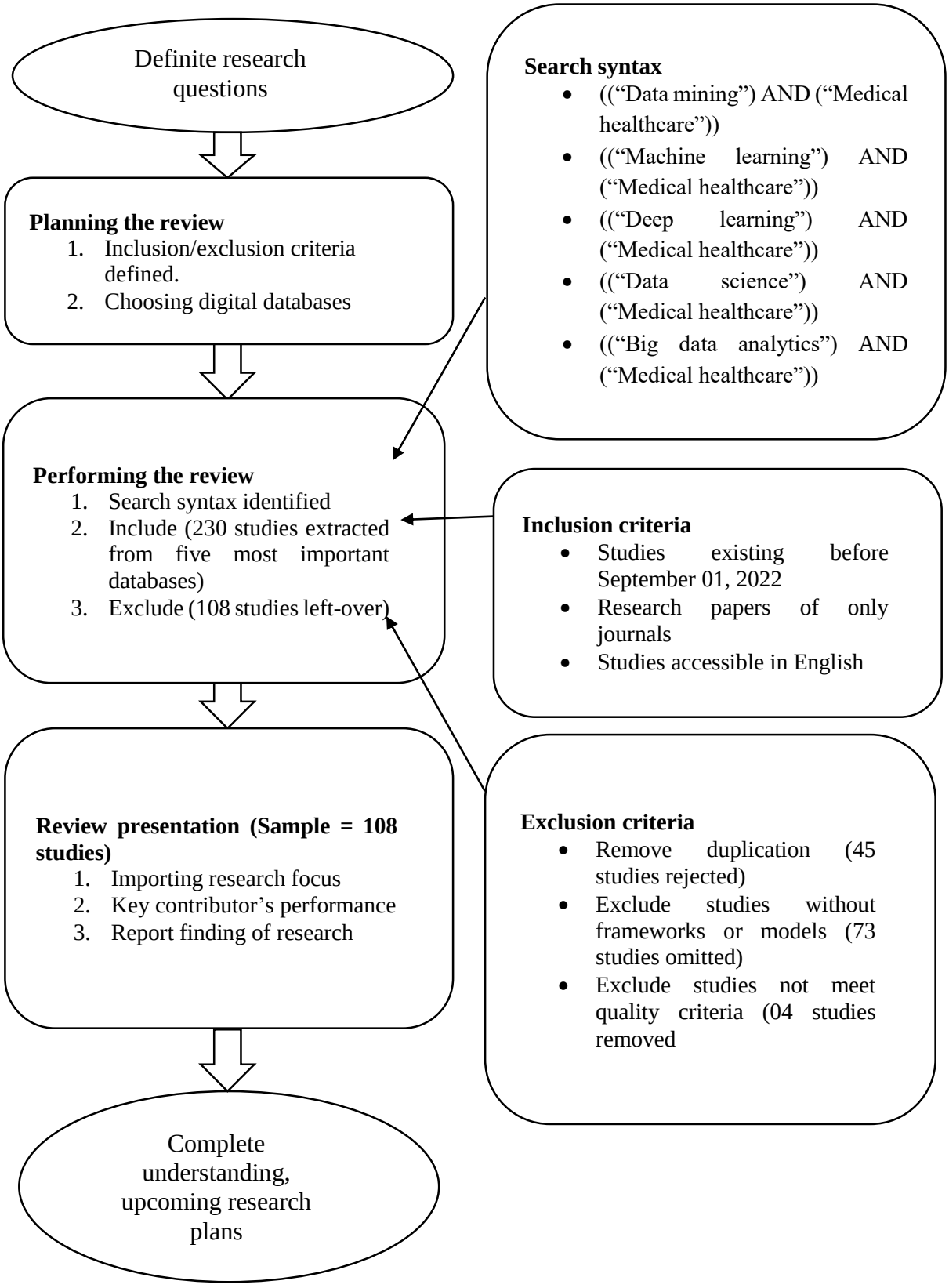

Figure 3.9: Flow chart of inclusion/exclusion studies in SLR

d. Discuss DS in medical healthcare.
e. Discuss BDA in medical healthcare.
f. Discuss DL in medical healthcare.
g. Discuss image-based methods.
h. Discuss signal-based methods.
i. Papers published in journals.
j. Papers in the domain are between”2015 to 2022”.
k. For history and simple information, the old studies may be considered.
l. The research questions must be addressed in existing studies.

3.7.7 Exclusion criteria

The limitations are listed below to make a proper exclusion criteria for searching.

a. Research papers not in the English language.
b. Redundant research papers.
c. Papers that are irrelevant to research questions.
d. Unauthorized publication and other websites without digital libraries.
e. The publication which is not published or unauthorized papers.
f. Papers in the domain are published before 2015.
g. Papers published in conferences.

3.7.8 List of journals

Table 3.6 list the most frequently referenced periodicals in the fields of software engineering and medicine. It is additional information that shows impact factor and listed journals are contained existing literature on the scope of digital libraries.

Table 3.6: List of included top studies publishing journals

S No	Journals	Impact factor	Data bases
1	Advanced research in computer engineering and technology	10.47	Sp
2	Expert system with applications	6.95	SD
3	NeuroImage	6.55	SD
4	IEEE access	4.34	IX
5	Medical informatics	4.04	SD
6	Electrical and computer engineering	3.81	SD
7	Healthcare engineering	2.68	Hindawi
8	Applied sciences	2.67	MD
9	Healthcare	2.64	MD
10	Scientific programming	2.14	Hindawi

3.8 Review conduction

When the protocol was recommended, then the further research is started in a proper way. However, in this section performed every stages created by a research protocol.

3.8.1 Identifying research questions

The key to any systematic review is clearly defining the research topics. The systematic review technique is driven by the review questions:

- The primary papers that answer the research questions must be found through the search method.
- The data that is needed to respond to the questions must be extracted during the data extraction procedure.
- The data must be synthesised during the data analysis process so that the questions can be addressed.

3.8.2 The pilot study

The research process is verified by a pilot study, and the research questions create a search string. To address the research questions, data analysis was done on all retrieved studies. It is essential

in the outset to assist the efforts in being directed in the proper direction. A pilot study is required to validate each aspect of the SLR review, including research keywords, strings, research questions, study selection, and data extraction. To address flaws in the data assembly and combination techniques, a study methodology must be implemented.

About nine primary studies was chosen as pilot studies, accounting for 10% of all research. Three studies from the quantitative-based technique, three from the image-based method, and three from the signal-based method was chosen. These investigations studied the entire method step by step and defined all of the parameters in great detail. The following are the research articles that was utilized as a pilot study: -

Table 3.7: The articles selected for a pilot study

Ref	Title	Data type
[28]	DM techniques in evaluating cardiac infection and documents of important features.	Quantitative-based
[32]	Improvement of a health big DM procedure by topic modeling.	
[34]	Rheumatoid arthritis identification in patients by adopting DM techniques.	
[51]	ML and anatomic signature based PET/MRI for brain attenuation correction.	Image-based
[52]	AD and mild cognitive impairment for automated classification used a particular MRI and DNN.	
[54]	A CNN application for health X-Ray image division is appropriate for small datasets.	
[74]	During the timed-up and go test, a trail was created to evaluate the limits required for sensing ECG and EEG associated illnesses.	Signal-based
[79]	EEG time-frequency research and SVM classifier are used to recognize emotional states.	
[86]	In remote medical applications, anomaly detection is built using ECG signal preparation and an SVM classifier.	

3.8.3 Study selection

The study selection criteria are to find primary studies that provide direct answers to the research question. Study selection criteria identifies in protocol definition and it is confirmed in search process. The research question also focused by inclusion and exclusion criteria. The piloted (sub-section 3.8.2) to ensure that constantly taken and categorized studies appropriately [17].

Table 3.8: Quantity of research papers and databases denoted with identification

Year	Databases	MD	SD	IX	Sp	OL
2015		-	2	1	1	2
2016		-	-	1	1	1
2017		-	-	1	1	3
2018		1	3	1	1	5
2019		-	3	1	1	4

2020	4	2	5	-	12
2021	8	8	12	6	14
2022	2	-	-	1	-
Total	**15**	**18**	**22**	**12**	**41**
Outline	Studies selected for the role of DM in medical healthcare				

3.8.4 Quality assessment (QA)

The quality is something adding into overall inclusion/exclusion standards, it is challenging to judge the "quality" of most important studies [17]. The checklist undermentioned below

a. The additional details included in inclusion/exclusion criteria.
b. To quality differences place positive impacts on study results.
c. Weighting is used in every investigation to assess the significance.
d. To elaborate the understanding of findings and regulate the strength of suggestions.
e. The further research enhanced by the guidelines.

3.8.5 Data extraction

The data extracted by selected studies based on important features that are support the research for attaining the goals. In this process to check the data extraction consistency therefore performed test-retest procedure on all extracted data. Table 3.9 shows the whole summary of databases.

Table 3.9: Databases summary

Database	Search queries	Total hits	Abstract read	Download
MDPI	(("Data mining") AND ("Medical healthcare")) (("Machine learning") AND ("Medical healthcare")) (("Deep learning") AND ("Medical healthcare")) (("Data science") AND ("Medical healthcare")) (("Big data analytics") AND ("Medical healthcare"))	399	25	12
Science direct		325	46	14
IEEE Xplore		2361	42	17
Springer		237	12	08
Other Digital Libraries		5502	105	38

3.8.6 Data synthesis

Data synthesis of primary investigations is used to summarize and collect the results. There are three forms of data synthesis: qualitative, quantitative, and descriptive (non-quantitative). However, it is feasible to combine quantitative and descriptive summaries at times. The information

gathered from research (diseases, data types, data sets, framework, sample sizes, classifiers, accuracies, and cross-validation characteristics) is tabulated in a way that is consistent with the review question. The tables in the study results emphasize the differences and similarities. These results are either homogeneous or heterogeneous, meaning they are consistent or inconsistent with one another. The influence of sources such as research type, sample size, and quality on the outcomes is listed. When a researcher is cautious about concluding a region as a whole from similar findings, they employ this strategy. Individual studies was first evaluated, and then the entire set of investigations was analyzed. Every research identifies, tabulates, and collects data on the mentioned relevant subjects. The qualitative (descriptive) nature of data synthesis is accomplished in this SLR [17].

3.9 Reporting the review

The final stage of SLR is to put down the review results that are useful to those who are interested. The final results are extremely crucial to convey since they represent SLR's most valuable discovery. Following the completion of the operation, the following is a summary of selected studies:-

Table 3.10: Summary of SLR results

Ref	Focusing area	Method used	Data type
[28]	Quantitative-based	A fast mutual information feature selection approach based on conditional mutual information (FCMIM).	Quantitative
[29]		Use the UCI stat log and Cleveland datasets to test hypothesis.	
[30]		Mutual information and recursive feature elimination.	
[31]		SLR of medical data and experiment for evaluation.	
[32]		The progress of a health big-DM procedure.	
[33]		DM techniques used a prototype.	
[34]		The optimum RA prognosis by prediction model.	
[35]		The Chinese material medical and the Chinese pharmacopoeia identified Chinese herb knowledge.	
[36]		Prediction of cardiac disease using three classification.	
[37]		The framework contains a RF and SVM algorithm enhanced by a slime mould algorithm (SMA).	
[38]		A method biology approach to the pathogenic procedure to classify essential standard as drug targets.	
[39]		A research plan of diabetes issues of DM techniques based on digital health record investigation.	

[40]		The biogeography-based optimization (BO) metaheuristic approach was used.	
[41]		This study involved 389 patients at Kermanshah's Imam Khomeini Hospital.	
[42]		The framework composed for data management to increase services within the South African medical conveniences.	
[43]		For feature mining and three classification models the principal component analysis (PCA) algorithm is used.	
[44]		AI used with NB and RF classification algorithms.	
[45]		The trend of demand patterns and demands in the dataset using an algorithm (Apriori association).	
[46]		The chi-square system was used to regulate the most significant features in identifying the COVID-19.	
[47]		Automatically extracting data from EHRs using text-mining can be used to identify trial 52 participants and to collect baseline information.	
[48]		ML practises based on probabilistic programming are used to discover pathway sorts that inspiration of recovery time.	
[49]	Image data	Classification of infections and organ/tissue subdivision, with a focus on AI and DL, which have a de-facto approach.	X-ray, CT
[50]		Systematically displays several unsupervised methods useful to health image analysis.	MRI, CT
[51]		A learning-based system for generating client CT maps in their inherent space from predicted T1-weighted MRI to eliminate brain PET variation.	PET/MRI
[52]		A DL predicting the specific analysis AD and mild cognitive impairment.	MRI
[53]		A fresh MRI model for transforming MCI to AD has been developed.	MRI, MCI
[54]		An edge resolution using ML that delivers robust and well-organized reasoning.	X-ray
[55]		AI-based and X-ray devices as fracture/non-fracture.	
[56]		On chest X-ray scans, the ML-based categorization of such mined depth element employing ResNet152 for COVID-19 and pneumonia cases.	
[57]		A pneumonia discovery method in a huge chest X-ray dataset.	
[58]		By analyzing chest X-ray data, EMC Net was able to categorize COVID-19 instances.	
[59]		DNN are ML methods across a variability of areas, from image examination to natural language handling.	MRI

[60]		A new handover category was introduced, which required candidates to report techniques evaluated on MRI machines outside of the training examples.	
[61]		Supervised AI methods for the assessment of health images need a curation procedure for data to optimally train, authenticate, and examination algorithms.	MRI, CT
[62]		Imaging sorts of postoperative complications in 40 cases were analyzed.	
[63]		Plan and build a wearable device that can determine the symptoms of attack throughout the elderly in a timely manner while walking	X-ray, CT
[64]		COVID-19 is differentiated using a dimension reduction approach that generates a collection of classification from CXR data for an effective ML algorithm.	X-ray
[65]		The DL model demands a huge amount of training samples associated with conventional methods.	
[66]		A multi-core computational model is used to fast-track the computational procedure.	
[67]		The detection of COVID-19 pneumonia diseased cases utilizing chest X-ray films has been classified using five pre-trained CNN methods.	
[68]		COVID-CT-MD will support the development of sophisticated ML and DNN approaches.	CT
[69]		On COVID-19 lung CT Diagnostic imaging, MUsculoskeletal RA radiographic (MURA) X-ray films, and high cholesterol, a multi-site Spatio-temporal split method was used.	X-ray, CT
[70]	Signals data	To detect heart medical issues and quantify intensity, the technique filters individuals' ECG and uses ML algorithms.	ECG
[71]		Arrhythmia identification utilizing ECG readings using traditional ensemble mining and DL approaches.	
[72]		Formulation of an NN-based approach for automatically detecting the association among prior knowledge symptoms in aged people and the characteristics estimated from different signals.	ECG, EEG
[73]		Create an accurate model for categorizing sleep phases using data collected from an ECG's heart rate variability (HRV).	ECG
[74]		Diseases associated with ECG and EEG information are recognized using sensors included in off-the-shelf smartphone gadgets linked toward a BITalino machine.	ECG, EEG
[75]		A CNN and a generalized minimax-concave (GMC) approach are used in the method.	ECG, EEG
[76]		Many illness datasets were classified using AI and the NB and RF classification methods.	ECG

[77]		To assess particular and generalized domain designs and pave way for real pressure assessment, supervised learning methods are implemented to categorize the generated highlighted clusters.	ECG, EEG
[78]		Applying k-NN models, the ML model based upon that MITDB arrhythmia collection (MIT-BIH Physionet) exhibited greater than 97 percent efficiency.	ECG
[79]		A tool for identifying feelings based on statistical data in the time series. The best factors was selected to use a box-and-whisker graph, which is then input into a Classification model for reviewing the DEAP dataset.	EEG
[80]		ML algorithms classify emotions in the arousal and valence dimensions compared the features in the time and frequency domain.	
[81]		An ensemble of P-1D-CNN models.	
[82]		Sleep-disordered breathing is among the six sleep disorders identified utilizing EEG data.	
[83]		CADFES, a new tool, was introduced.	
[84]		Drowsiness recognition system associated with is physiological cues.	
[85]		To investigate stress utilizing various EEG data pathways.	
[86]		To distinguish between normal and sick participants, ECG condition characterized and SVM-based cardiac arrhythmia wave prediction is used.	ECG
[87]		A look at how ECGs are categorized into different forms of arrhythmias.	
[88]		Develop and deploy a wearable health system for the detection of the signs or symptoms of stroke inside the aged in real-time while walking.	EEG
[89]		There is scientific evidence that diseases may be reliably diagnosed utilizing implanted EEG sensors and non-EEG equipment.	
[90]	Deep learning, Data science, Big data analytics	Recent Deep Learning Techniques , Challenges and Its Applications for Medical Healthcare System : A Review	DL
[91]		Benchmarking deep learning models on large healthcare datasets.	
[92]		Multi-View Deep Learning Framework for Predicting Patient Expenditure in Healthcare.	
[93]		COVID-19 Detection Using Deep Learning Algorithm on Chest X-ray.	
[94]		A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram.	DS
[95]		International Journal of Medical Informatics Healthcare professionals' acts of correcting health misinformation on social media	
[96]		A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming.	

[97]		IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning.	
[98]		Real-time Medical Emergency Response System : Exploiting IoT and Big Data for Public Health.	BDA
[99]		A Security Management Framework for Big Data in Smart Healthcare.	
[100]		Big Data , Big Knowledge : Big Data for Personalized Healthcare.	
[101]		Real World — Big Data Analytics in Healthcare.	
[102]		A systematic literature review of data science , data analytics and machine learning applied to healthcare engineering systems systems.	DS
[103]		Role of machine learning in medical research : A survey.	
[104]		Machine Learning in Healthcare Data Analysis : A Survey.	
[105]		A SLR of Medical Image Analysis Using Deep Learning.	DL
[106]		Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review	BDA
[107]		Data mining and predictive analytics applications for the delivery of healthcare services : a systematic literature.	
[108]		Big Data Analytics in Healthcare — A Systematic Literature Review and Roadmap for Practical Implementation.	

CHAPTER 4

ANALYSIS AND RESULTS

4.1 Overview

The comparison in this chapter has been made between many existing studies that are the results of data collected from the SLR. Existing characteristics, methodologies, and approaches are used to map the comparison. Datasets, DM approaches, cross-validation procedures, sample size, data type, classifiers, and accuracies are among the characteristics. After that, present the review's findings. At the end of the chapter, there is a discussion of the presented outcomes.

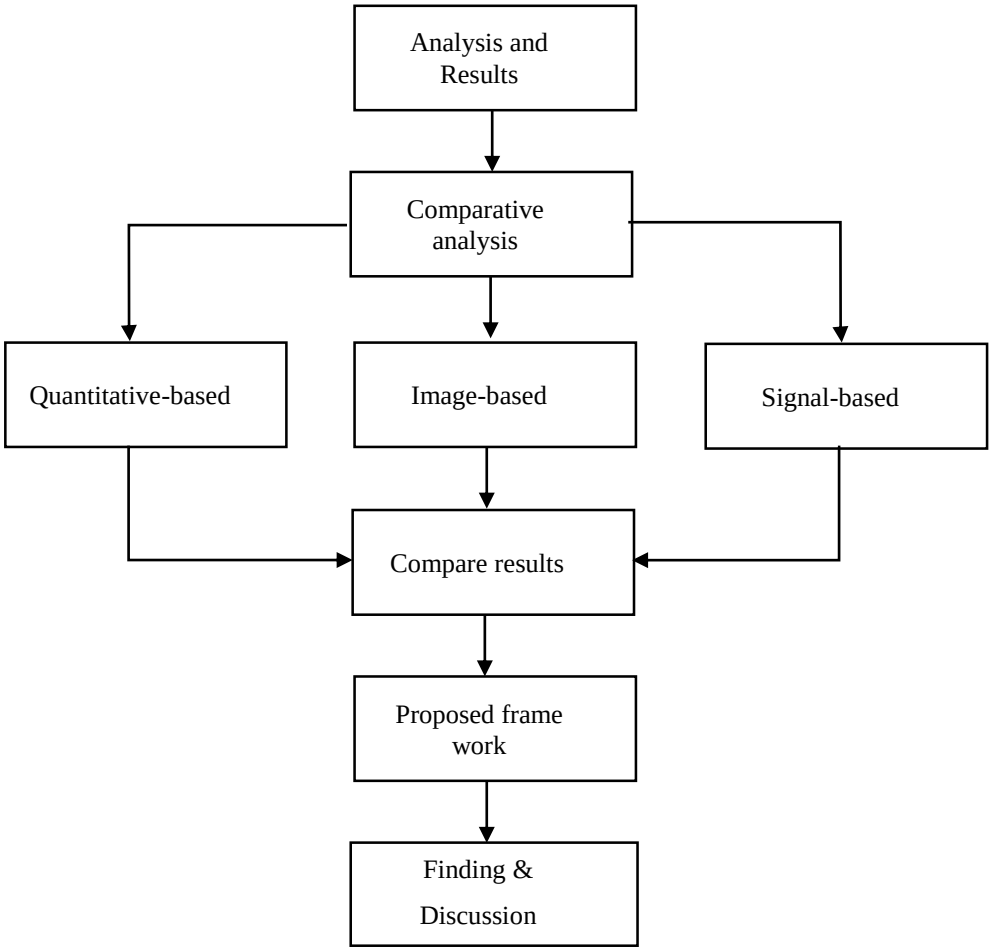

Figure 4.1: Layout of analysis and results chapter

4.2 Comparative analysis of data mining in medical healthcare

In the previous studies, old-style skills and to store, process and manage the data warehouses was used by data scientist. However, the huge volume of data revolution in medical cannot be controlled using simple database organisms, tools, and methods. Nowadays, various advanced skills with high calculating power and storing capacity have established to discourse the low routine and difficulty of old-style systems. Big data equipment denoted to as progressive knowledge that have the great computing power and logical ability to development large sizes of data composed from various bases to extract understanding from it. DM systems cover an extensive range of fields like statistical examination, ML and image investigation [29].

This SLR on the presentation of DM in medical is dedicated on three major methods i.e. quantitative-based methods, image-based methods and signals-based methods. These three methods have contained data in form of dataset records, images and signals like statistical tables, x-ray, CT, MRI, fMRI, ECG and EEG form respectively. Now are doing a deep analysis of these abovementioned methods.

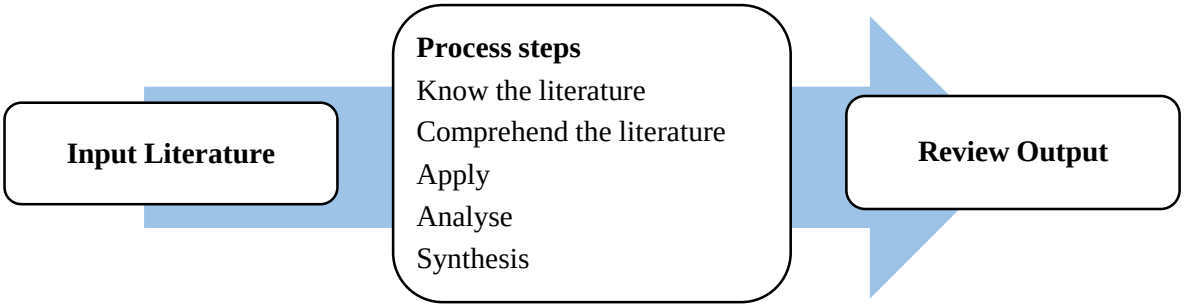

Figure 4.2: Analysis and result process

4.2.1 Examination of quantitative-based method

JIAN PING LI et al. [28] used the Cleveland heart infection dataset in this existing study. A well-organized and exact system for detecting heart infection, as well as an approach using ML techniques. SVM, LR, ANN, KNN, NB, and DT are some of the classification algorithms that are used in the procedure. Leave one subject out (LOSO) is a validation approach in which one sample is saved as test data while outstanding individuals are used to train the experiment. The one subject chosen for model cross-validation was employed for understanding of the best performance of model evaluation and for super constraint regulation. The investigational outcomes show that the existing feature collection algorithm, fast conditional mutual information (FCMIM) is practicable with a classifier SVM for planning a high level quick system to recognize heart infection. The accurateness of SVM with the collected feature nominated algorithm, fast conditional mutual information (FCMIM) was 92.37% which was best as associated to previously methods.

Mohammad Shafenoor Amin et al. [29] employed a prediction system and seven classification algorithms, including KNN, DT, NB, LR, SVM, NN, and vote, to build a grouping of characteristics (combine with NB and LR). The heart infection prediction model created utilising the identified essential characteristics and the DM approach (i.e. Vote) achieves an accuracy value of 87.4 percent, according to trial results. The experiment was initially carried out using data from UCI Cleveland to identify the key characteristics and the top three DM techniques. An additional experiment using the UCI Statlog dataset was utilised to analyse the findings. Vote, NB, and SVM are three DM approaches that have shown to be the most accurate in prediction in previous study. The assessment outcomes sure that the nine designated features are noteworthy. Moreover, among the methods, Vote has outstanding then other two methods. So the top prediction system was produced by the nine significant features and the Vote method.

Chest discomfort is a common complaint in emergency rooms, according to Chieh-Chen Wu et al. [30]. An ANN technique was used to conduct a trial research to regulate the prospective risk of major adverse cardiac events in Taiwan. Five biomarkers were choose from a collection of 37 participant characteristics using an ML-based feature selection technique. A comprehensive and well-managed risk categorization system has been developed. The whole model has a sensitivity of 94.8 percent, a specificity of 54.6 percent, a true positive rate of 22.8 percent, as well as a diagnostic accuracy of 98.7%.

Table 4.1: Compression of quantitative-based method

Author, Ref, Year	Method	Disease	Data Set	Sample Size	Techniques	Classifiers	Cross-validation	Accuracy
Jian ping li et al. [28] 2020	Quantitative based	Heart	Cleveland Heart disease	303	Fast Conditional Mutual Information	SVM	LOSO	92.3%
Mohammad Shafenoor Amin et al. [29] 2018	Quantitative based	Heart	UCI Cleveland+ UCI stat log	303	Vote (combine with NB and LR)	NB SVM	Confusion matrix	87.4%
Chieh-Chen wu et al. [30] 2020	Quantitative based	Chest pain	The cardiac event in Taiwan	1175	ML-based feature selection	ANN	Risk score model	94.8%
Chang woo song [32] 2017	Quantitative based	BP Diabeties	Health insurance review & Assessment service	1000	The bag of words technique & the latent dirichlet allocation method	DM Classification	Medical DM & accuracy of the document list recommendation	75%
Chien-Ting Wu et al. [34] 2020	Quantitative based	Rheumatoid arthritis	Medical centers in southern Taiwan	3486	Using DM methods to investigate the ESR value.	DT	The dependent variable in WEKA	95.15%
Ping Xia et al. [35] 2020	Quantitative based	Kidney	Jiangsu province hospital	166	Clinical circumstances, combination of association regulation, cluster examination and complex system examination.	Apriori Algorithm	Complex network analysis	95.8%
Sujata Joshi [36] 2020	Quantitative based	Heart	UCI learning repository	303	Ranker method from WEKA	DT NB KNN	Gain ratio attributes Eval	92% 84% 96%
Peiliang wu [37] 2021	Quantitative based	COVID-19	Box plot of 28 indexe statistics analysis	26	Comparative experiment RF-SMA-SVM method, (Slime mould algorithm)	RF SVM	10-fold	92%
Hindreen Rashid Abdulqadir [38] 2021	Quantitative based	Diabeties	UCI Pima Indian diabetes registry	768	Biology approach to the pathogenic pathway	RF	10-fold	75%
Touraj Ahmadi jouybari [41] 2021	Quantitative based	Burn	Imam Khomeini hospital of Kermanshah city	389	Cross-sectional study including demographic information, geographical and burn information.	DT	Logistic regression	87%

Rheumatoid arthritis (RA) is indeed a rational chronic inflammatory infection, according to Chien-Ting Wu et al. [34] in previous research. In 2 health centres in southern Taiwan, detailed records of outdoor patients were used as a study sample. The accuracy rate of the estimated model using the LR, SVM, and DT models was 79.2 percent, 78.2 percent, and 90.9 percent, respectively, according to the results. The current study compares the average accuracy of the two datasets using the LGR (Basic Linear), SVM (SMO), and DT single classifiers, finding that DT is more accurate than LGR.

Ping Xia et al. [35] used a comprehensive approach based on medical hospital cases, with the inclusion of complicated network analysis, cluster investigation, and association techniques, to determine effective herbal prescriptions for CKD therapy. The author examines a patient trial from Jiangsu Province Hospital of Chinese Medicine retrospectively. In all of the treatments, the Apriori algorithm examined the herb association techniques. Support and confidence levels are the two most important restrictions. The right association processes received a 51.51 percent level of support, with a 95.88 percent level of trust.

4.2.2 Examination of image-based method

In the existing study, Parnian Afshar et al. [68] the COVID-CT-MD dataset, which includes patient-level, lobe-level, and slice-level labeling, has the potential to aid COVID-19 research. COVID-CT-MD, in particular, can aid in the establishment of incremental ML and DNN-based resolutions. A volumetric breast CT scan is included in the dataset. COVID-19 circumstances are collected at Babak imaging centre in Tehran, Iran, from in 2020, whereas community-acquired pneumonia (CAP) and normal circumstances are collected in 2020, respectively. Gender accuracy is 108 out of 171 for males and 63 out of 171 for females, with ages ranging from 37 to 66. From 60 CAP instances, 35 was male and 25 was female, with ages ranging from 36 to 79.Normal 76 cases have 40 male and 36 female, their age between 29 to 58 years.

In another study, Joseph Bullock et al. [54] developed a CNN for healthcare electromagnetic radiation (X-ray) photograph subdivision with an accuracy of 92 percent, an F1 score of 92 percent, and an AUC of 98 percent, outperforming traditional image processing techniques like clustering and entropy-based methods. The data was collected from two foundations at (IBEX Innovations Ltd 69)

Method (e.g. of feet, knees, and phantom heads, as well as 81 conventional X-Ray images of various body and phantom body components.

Rahul Kumar et al. [56] observed that the ML-based categorization of extracted deeper features by ResNet152 combined COVID-19 and pneumonia patients on chest X-ray images resulted from incorrect guesses of COVID-19 utilizing chest X-Ray images. A total of 5840 images was utilized in the study, with 5216 images used for training (1341 for normal class & 3875 for pneumonia class) and 624 (234 for normal class with 390 for pneumonia class) in testing. The data sets of COVID-19 versus normal patients was matched using a technique known as synthetic minority oversampling (SMOTE). Utilizing XG Boost prognostic classifiers, the system achieves an accuracy of 97.3 percent on RF and 97.7% on RF. Two datasets was employed in this research: chest x-ray pneumonia as well as the COVED-19 public collection on the knowledge graph.

Another research, conducted by Prottoy Saha et al. [58], used an automated discovery system called EMC Net to identify COVID-19 patients by analysing chest X-ray images. EMC Net's collection had a total of 4600 images. There was three sets of images in the dataset: a training set with 3220 images, a validation set with 920 images, and a test set with 460 images. For the detection of COVID-19, a CNN was built using binary ML classifiers (RF, SVM, DT, and Ada Boost). EMC Net uses CNN to extract high-level sorting from X-ray images. With 98.91 percent accuracy, 100 percent exactness, 97.82 percent memory, and 98.89 percent F1-score, an EMC Net grouping of classifiers exhibited better presentation.

Table 4.2: Compression of image-based method

Author, Ref, Year	Method	Disease	Data Set	Sample Size	Techniques	Classifiers	Cross-validation	Accuracy
Xiaofeng yang [51] 2018	PET & MRI	Brain Condition	Cohort of 17 patients randomly selected	17	CT img served as regression targets of paired MR images	RF	LOSO	97.0%
Silvia basaia [52] 2018	MRI	Alzheimer's disease (AD)	AD neuroimaging initiative (ADNI)	1409	DL algo predict AD & mild cognitive impairment will convert to AD(c-MCI) based on single cross-sectional brain structural MRI scan	CNN	10-fold	98%
Elaheh Moradi [53] 2015	MRI	Alzheimer's disease	AD neuroimaging initiative (ADNI)	1500	MRI biomarker of MCI-to-AD conversion	RF	10-fold	76.6%
Joseph Bullock et al. [54] 2019	X-Ray	Bone and soft tissue	IBEX Innovations Ltd	81	Clustering & entropy, classical image processing, X Net	CNN	Images without augmentation	92%
Fatih Uysal [55] 2021	X-Ray	Shoulder fractures	Musculoskeletal radiographs	8942	Ensemble learning models	CNN	Confusion matrix with AUC	84%
Rahul Kumar et al. [56] 2020	X-Ray	COVID-19	Covid-19 dataset from Italy	5840	ResNet152	RF	Confusion matrix with AUC	97.3%
Sini Tang [57] 2021	X-Ray	Pneumonia	Chest X-ray 14	3110	Shapley algorithm	CNN	LOSO	51%
Prottoy Saha et al. [58] 2020	X-Ray	COVID-19	Github repository developed by Cohen	4600	EMC Net	CNN	10-fold	98.9%
Elaziz et al. [66] 2020	X-Ray	COVID-19	Team of research Qatar University	216	Mobile Net Model	KNN CNN	10-fold	96% & 98%
Parnian Afshar et al. [68] 2020	CT Scan	COVID-19	Babak imaging center, Iran	307	Lob-level, slice-level, patient-level Labels	DNN	Quality control	82%

4.2.3 Examination of signal-based method

Turker tuncer et al. [71] used a discrete wavelet transform (DWT) in combination with a novel 1-dimensional hexadecimal local pattern (1D-HLP) approach to find arrhythmias automatically. Using the MIT-BIH Arrhythmia ECG dataset, they was able to classify 17 arrhythmia modules with a 95.0 percent classification precision. The extracted features are sent into the KNN Classifier, where 1NN is the simple kind and k is one.

In another existing study, Muhammad Wasimuddin et al. [76] offered an evaluation of 168 research that is a stages-based approach for ECG signal examination in previous work. At each level of the assessment, the system defines the old-time/frequency realm with advanced ML approaches as indicated in the available literature. The author recorded clinical ECG data and acquired a real-time ECG signal. It was possible to use ML to detect R-peaks and QRS complexes. In a recently published study, the deep learning strategy produced more efficient finding and classification results. In the form of a table, the author summarises a selection of DL approaches for ECG evaluation that has recently been described in the literature. According to the study, most researchers used the MIT database to develop an ECG examination and classification based on one-dimensional ECG data.

In another existing study, Fabian Parsia George et al. [79] suggested an emotion finding algorithm based on time-frequency domain statistics sorting in another previously published paper. A dataset for emotion analysis through physiological data (DEAP) dataset, wherein participants are 32 and are separated by gender and age, is utilized to find the best characteristics, which are then input to an SVM for training and validation. Two types of emotions was identified using pre-arranged EEG signals out from the DEAP dataset: valence and arousal. For the tested dataset, the results show a 92.36% accuracy.

In another existing study, Ihsan Ullah et al.[81] developed the approach using the university of Bonn dataset, which is a standard dataset; the accuracy is 99.10 percent in practically all cases of epilepsy identification. The P-1D- CNN (pyramidal one-dimensional deep CNN) system is suitable for epilepsy detection and is widely used in creating robust skilled systems for other illnesses. The approach is a combination of memory-capable and basic P-1D-CNN methods that takes an EEG signal as input, feeds it via various P-1D-CNN methods, and finally tempers the findings with a public vote. Each set (A to E) has 100 one-channel occurrences.

Table 4.3: Compression of signal-based method

Author, Ref, Year	Method	Disease	Data Set	Sample	Techniques	Classifiers	Cross-validation	Accuracy
Turker tuncer et al. [71] 2019	ECG	Arrhythmia	MIT-BIH	1000	One dimensional hexadecimal	KNN	Classical ensemble & DL	95.0%
Nico surantha [73] 202	ECG	Sleep stage	MITBPD	18	Categorizing sleep phases by sorts of heart rate variability mined from ECG.	SVM	70% test data and 30% training data was used for validation	72.2%
Vasco Ponciano [74] 2020	ECG & EEG	Heart	Biological signal, time up & go test	40	Multiplayer perceptron method, implemented with the WEKA software	NN	Two-way analysis of variance test	93%
Muhammad wasimuddin et al. [76] 2020	ECG	Heart	MIT database	107	Stage based process model, classification	NN	2-D image classification	99.5 %
Fabian Parsia George et al. [79] 2019	ECG	Emotional States	Database of emotional analysis using physiological signals.	32	Emotion detection method based on time frequency domain statistical features	SVM	10-fold cross-validation	92%
Ihsan Ullah et al. [81] 2018	EEG	Epilepsy detection	University of Bonn	4097	Pyremidal one dimensional deep CNN model	CNN	10-fold cross-validation	99.1 %
Manish sharma [82] 2021	EEG	Sleep disorder	Sleep disorder center of the ospedale maggiore of Parma,	108	A method using EEG signals for the automated identification of six sleep disorders	SVM	10-fold cross-validation	91.3 %
Shivarudhrappa Raghu et al. [83] 2018	EEG	Epilepsy detection	Bern-Barcelona database	7500	Computerized automated detection of focal epileptic seizures	SVM	10-fold cross-validation	96.1%
Lan-lan chen [84] 2015	EEG	Drowsiness	Nihon koden EEG 2110	2048	A system of drowsiness detection using physiological	SVM	10-fold cross-validation	94.7%
C. venkatesan [86] 2018	ECG	Arrhythmic beat	MIT-BIH Arrhythmia DB	200	Arrhythmic beat classification and abnormality detection	SVM	Other ML classifiers	96%

In another existing study, Shivarudhrappa Raghu et al. [83] developed a technique called computerized automated detection of focal epileptic seizures (CADFES) in a previous study. 41.66 hours of EEG data from the Bern-Barcelona study was used in total. The dataset consisted of 3750 focal and 3750 non-focal EEG records from 5 pharmacoresistant time-based lobe epilepsy patients. SVM, K-NN, RF, and ada-Boost classifiers was used to assess the process's visual presentation. The SVM classifier produced the best results with sensitivity, specificity, accuracy, positive predictive rate, negative predictive rate, and AUC of 97.6%, 94.4 percent, 96.1 percent, 92.9 percent, 98.8%, and 96 percent, respectively.

4.3 Comparative results of Data Mining in medical healthcare

The following table summarises data from the above-mentioned individual technique examinations for the reader's convenience. Authors, years, diseases, data types, sample size, data sets, methodologies, classifiers, cross-validation, and accuracies are all included in table 4.4.

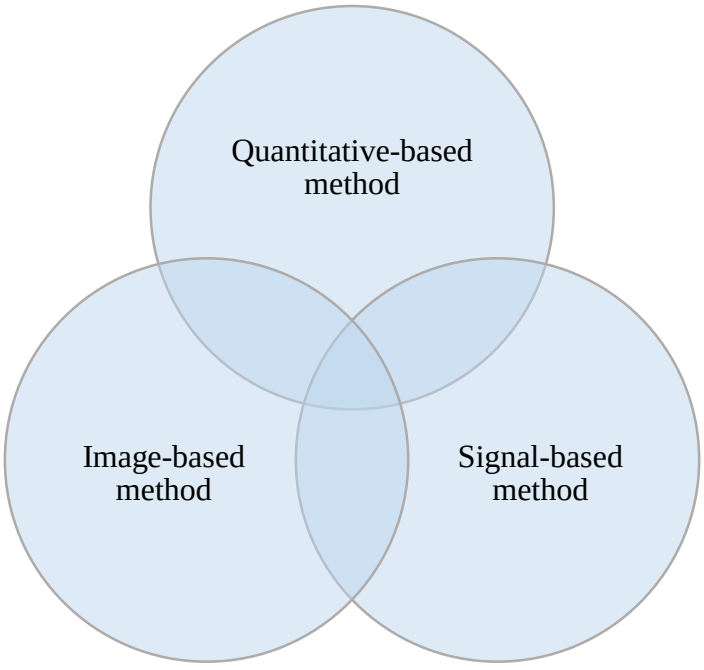

Figure 4.3: Comparison of three methods

Table 4.4: Comparison of components used in previous studies

Articles	Years	Methods	Diseases	Data Sets	Sample Size	Techniques	Classifiers	Cross-validation	Accuracy
[28] [29] [30] [34] [35]	2018 to 2020	Quantitative based datasets	Heart, chest pain, kidney, COVID-19	Cleveland Heart disease, UCI Cleveland+ UCI statlog, The cardiac event in Taiwan and Jiangsu province hospital	303, 1175, 714, 166 and 26	Fast Conditional Mutual Information, Vote (hybrid tech with NB and LR) ML-based feature selection Clinical circumstances, the combination of association regulation, cluster investigation, and complex system examination	SVM, NB, ANN, DT, Apriori Algorithm	LOSO, confusion matrix, risk score model, dependent variable. K-fold, complex network analysis.	92.3%, 87.4%, 94.8%, 54.6%, 98.1%, 51.5%, 95.8%, respectively
[68] [54] [56] [58] [66]	2019 to 2020	CT Scan, MRI and X-Ray	COVID-19, Bone and soft tissues, AD.	Babak imaging center, Tehran, Iran IBEX Innovations Ltd Covid-19 public dataset from Italy Github repository developed by Cohen (Author) Team of research Qatar University Doha, Qatar	307, 81, 5840, 4600, 216, 1500, 1409.	Lob-level, slice-level, patient-level labels Clustering & entropy, classical image processing, XNet ResNet152, ML EMcNet MobileNet Model	DNN, CNN, RF, KNN	Confusion matrix, AUC 10-fold.	COVID 7.0%-82% CAP 7.8%-56.8% 92% 97.3% & 97.7% 98.9% 96% & 98%
[71] [76] [79] [81] [83]	2018 to 2020	ECG and EEG	Arrhythmia, emotional state, epilepsy detection, Drowsiness, Sleep stage.	MIT-BIH Arrhythmia ECG dataset, MIT database, Emotional analysis using physiological signals The University of Bonn, Bern-Barcelona database	1000, 107, 32, 4097, 7500, 2048, 18.	One dimensional Hexadecimal, Stage based process model, Classification Statistical features Pyremidal one dimensional deep CNN model Computerized automated detection of focal epileptic seizures	KNN, NN, SVM, CNN	Classical ensemble, 2-D image classification, 10-fold.	92.3% 99.1% 96.1%

4.3.1 Common components

Heart disease is a regular finding in this suggested study quantitative-based strategy. In the most available research, the classifiers SVM and DT are repeated with 90% or higher accuracy. The sample size ranges from 300 to 1000 people. For cross-validation, the K-fold and confusion matrix are commonly utilized. These investigations was completed in 2020. The X-ray and MRI data format is commonly utilized in image-based methods, and the CNN classifier is employed in numerous investigations. In 2020, the average number of studies was reported. COVID-19, a regularly used test for Alzheimer's disease, has sample sizes of up to 1500 people and uses 10-fold cross-validation to evaluate the findings. The average success rate of image-based approaches is 85 percent. The ECG/EEG data type and SVM classifier reflect an average of 94 percent accuracy in the signal-based technique. The sample sizes range from 18 to 200 people. From 2018 through 2021, arrhythmia, epilepsy detection, sleep disorders, and cardiac illnesses were studied. Two research used the MIT data set, and most studies employed the repeated 10-fold cross-validation procedure to examine the results.

X-rays are contained 2-D radiations images and take 10 minutes max time for test process. The cost in between $260-$460 and diagnose the diseases bones cancer, pneumonia, tumours & infections. But on the other hand CT scan contained with 3-D, 360 degree images and time required for test process is 05 minutes only. Cost of CT scan is too high $270-$5000 and diseases diagnosed are organ, soft tissues & heart disease etc. MRI have 3-D magnetic & radio waves and 30 minutes time for test is required. The cost of MRI is $1600-$8400 and diseases tested are brain, muscles, neck etc.

ECG is signals based technique developed for heart electro plus during cardiac cycle and required only 05-10 minutes time for test. Its cost is less in every region that is $175-$299 and applicable for heart disease. EEG is the electrical activity of brain and 20-40 minutes time required for test. The cost of EEG is $200-$3000 which higher than ECG and brain diseases diagnosis, emotions analysis done by EEG. The last is quantitative datasets that are tabular & statistical data which is advised by medical professionals. Thousands of datasets are freely available online and time for pre-process capable for usage is required. Common diseases with doctor’s observations available in quantitative datasets.

4.3.2 Different components

Chest pain, kidney, burn, coronavirus, and rheumatoid arthritis disorders are all included in this suggested study's quantitative-based strategy. Datasets and techniques were gathered from a variety of places. The classifiers employed include the apriori algorithm, KNN, and ANN. Their precision is similarly around an average of 90%. Cross-validation approaches such as leave one subject out (LOSO), complex network analysis, and risk resource model were employed by research authors in these previous investigations. The dataset and methodologies for the quantitative-based approach are applied in many ways. In 2018 and 2021, certain past research will appear differently way. The sample sizes ranged from less than 300 to over 1000. Pneumonia, brain, bone, soft tissues, and shoulder fractures have all been studied using the image-based technique.

The methodology and datasets used in each study differ. Some studies, such as 5840 and 8942, have a big data sample. The classifiers DNN, KNN, and XG boost occur in diverse ways. Table 4.4 lists the datasets and approaches that were employed in all methods. In 2015, a few research on signal-based approaches appeared, and sickness emotional states and tiredness were examined to test and obtain findings. Large sample sizes, such as 1000, 2000, 4000, and 7500, are used in some studies. Different datasets and methodologies were employed by each author. A few research employed the CNN classifier, and for cross-validation, several approaches such as classical ensemble and analysis of variance test were used.

4.3.3 Best components and features

According to accuracies and illness prediction, SVM is the best classifier employed in this suggested study for the quantitative-based strategy achieved 92% results. The WEKA tool's ranker approach yields the most accurate 96% results. However, the average outcome of all quantitative-based investigations is 90%. The image-based technique, which combined x-ray plus MRI data with CNN, produced the greatest results with 98 percent accuracy but the overall average accuracy was 85%. ECG and EEG data is typed using an SVM classifier in the signal-based technique, with the best results reaching 96% but the average accuracy of all studies was 94%. To evaluate the results 10-fold was the best technique in the signal-based method.

4.3.4 The proposed framework

After examining the SLR results, this study provided the framework. This is the study's key contribution. In this suggested study, three different methodologies are discussed: quantitative, image-based, and signal-based. Compare the three ways of DM stated above, as well as the performance of classifiers and the outcomes.

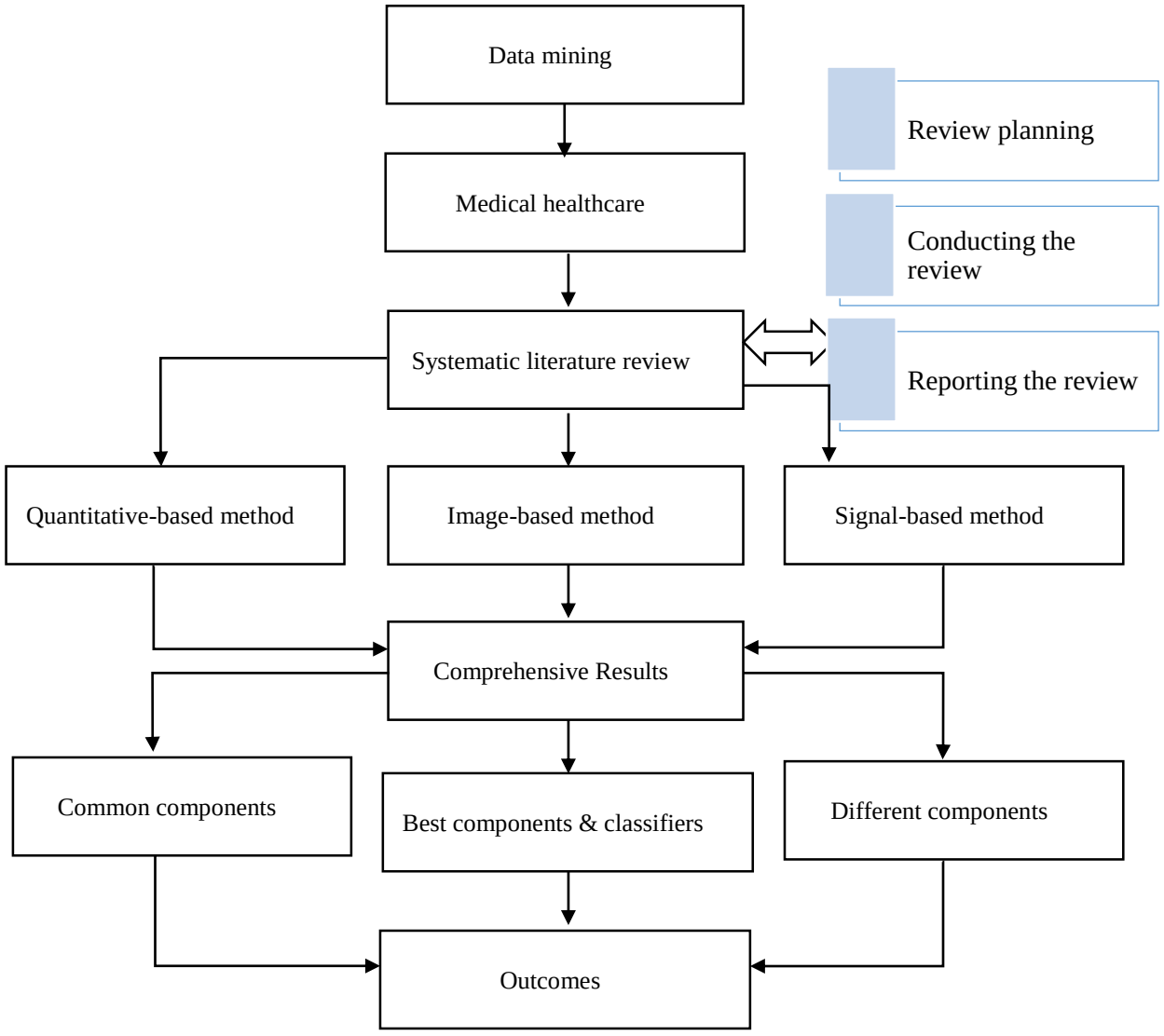

Figure 4.4: The proposed frame work

4.4 Findings of research

The findings of this study were analyzed using NVivo, the most widely used software for qualitative data analysis. DM is built on sophisticated algorithms that enable data segmentation to find multiple trends and patterns, spot outliers and calculate the probability that specific events will occur [1]. The format of raw data, which can be analog or digital, mostly depends on where the data came from [3]. To compete in the market and outperform the competition, businesses must stay up to date with the most recent DM trends [6]. Applications are explored in DM trends: Scalable and interactive DM techniques can handle application-specific challenges. DM integration with cloud computing systems, database systems, data warehouse technologies, social and information networks, spatiotemporal, moving objects, and cyber-physical systems, multimedia, text, and online data, biological and biomedical data mining, Real-time data stream mining, distributed data mining, and audio-visual data mining [10].

Modeling is the process employed to carry out these achievements. It is the process of creating a set of instances or a mathematical relationship based on data from circumstances in which the solution is known and then applying the model to additional circumstances in which the solution is unknown [13]. Although techniques have been used for centuries, the data storage and communication capabilities needed to collect and store enormous amounts of data, as well as the computational power to automate modeling techniques so they can be used directly on the data, have only recently become available [23]. The mathematical procedures and approaches employed in DM are frequently well-known [33]. The use of those methods to solve common business issues is novel, made possible by the greater accessibility of data and the low cost of storage and processing power. Additionally, tools that business specialists may easily utilize have become available as a result of the adoption of graphical user interfaces [102]

ANN, which are non-linear predictive models that learn through practice and have a topology similar to biological neural networks, is one of the techniques used for DM. DT structures are formed like trees and symbolize collections of choices. These choices produce the categorization rules for a dataset. The process of deriving usable if-then rules from data using statistical significance is known as rule induction. The principles of genetic combination, mutation, and natural selection are the foundations for the optimization methods known as genetic algorithms. A categorization method called nearest neighbor assigns each record in a historical database a category based on the records

that are most similar to it [107]. A ML application is a computer program that develops knowledge without human programming. A pattern in the data is found by the computer algorithm, which then forecasts the apparent result. AI can be attained by ML [11]. Software with hard coding is fed to machines. To complete tasks, the program provides detailed instructions. An algorithm for a machine is trained to complete tasks. Data is fed into the algorithm during the process, which enables the algorithm to grow and govern itself. The capability to identify an object in an image is an example of ML. The system is said to have "learned" when it detects an image with a high level of accuracy. Learning machines are a technique that PayPal utilizes to protect its systems from fraud [22].

High dimensionality is the cause of the ML trend. The program enables machines to think and make decisions automatically in the same ways that people do. Machines are capable of strategic planning and creativity. Advancements in technology ML is a development in technology, not a fad or hype. We can handle the most important problems of the present with ML. reduces overload at the moment; data are plentiful. A rising amount of data is being gathered and kept through emails, social networks, blogs, webinars, RSS, and podcasts. It's challenging to keep track of useful data. Finding your information is simple thanks to the advent of ML. The appropriate tools used by ML can locate your data in a structured way. Large volumes of disparate data from people, places, organisations, and questions are used in complicated situations that ML can manage. ML saves time and money by performing a lot of computation quickly and affordably. Creating reports on client activity and email tracking, for instance [28].

A wide, multidisciplinary notion called "digital health" or "digital healthcare" includes ideas from the nexus of technology and healthcare [15]. Software, hardware, and services are all part of digital health, which applies digital transformation in the healthcare industry [42]. Mobile health apps, electronic health records, electronic medical records, wearable technology, telehealth, telemedicine, as well as personalized medicine are all included under the term "digital health" [48]. Patients, healthcare professionals, researchers, application developers, medical device makers, and distributors are some of the stakeholders in the field of digital health. Today's healthcare system is becoming more and more dependent on digital technology [95]. Deep learning is a class of ML that uses deep neural networks, which are very large artificial neural networks. Neural networks are systems that can learn from data [5]. They don't need to be told how to do anything to do better. As a result, DL is seen as a significant player in the artificial intelligence industry [21]. Recent developments in DL include leveraging larger datasets and more complex structures, as well as integrating interactions

between various neural network types and other AI technologies like decision trees and natural language processing [26]. To make a diagnosis, DL models can decipher medical pictures like X-rays, MRI scans, CT scans, etc. In medical imaging, the algorithms may identify any risk and highlight irregularities. The detection of malignancy makes considerable use of DL [103]. DL has made it possible to perform computer-aided disease diagnosis and detection [104]. Through the process of medical imaging, it is frequently utilized for medical research, medication development, and identification of fatal diseases including cancer and diabetic retinopathy [105].

To get better insights and empower healthcare practitioners to make well-informed decisions, data analytics in the healthcare sector automates the gathering, processing, and analysis of complex healthcare data [9]. With better scheduling and staffing, data analytics in clinical settings aims to decrease patient wait times, give patients more options for scheduling appointments and receiving care [18], and lower readmission rates by using population health data to identify the patients who are most at risk [20]. Five basic categories of analytics may be distinguished: descriptive, diagnostic, predictive, prescriptive, and discovery analytics. Each of these categories plays a specific function in enhancing healthcare [27]. Predictive analytics are used in the healthcare industry, for example, to reduce readmissions, manage population health, improve cyber security, boost patient engagement and outreach, expedite insurance claim submission, forecast appointment no-shows, and prevent suicide attempts [100].

Data science aids in the detection of scanned images to identify human body flaws and assist clinicians in developing an efficient treatment plan [8]. A few examples of these medical picture tests are X-rays, MRI, and CT scan. There are several uses for DS in the medical field. The medical and healthcare sectors have made extensive use of DS to improve patient lifestyles and identify ailments early on [25]. In addition, because of developments in medical image processing, clinicians can now detect small malignancies that were previously difficult to detect. Consequently, DS has significantly changed healthcare and the medical sector [108]. In-depth images of your internal organs can be obtained using contemporary imaging techniques like X-rays, ultrasonography, CT scans, and MRI [49]. Medical professionals can identify medical issues and track changes in the body using procedures like MRIs, CT scans, x-rays, and others. The healthcare sector requires additional technologists to fill critical jobs as the uses of medical imaging technologies are developing quickly [50]. Rescaling photos (Digital Zoom), lighting correction, edge detection, mathematical morphology, and evaluation and ranking of segmentation methods are a few examples of image

processing [51]. The process of removing useful information from medical photographs involves frequently applying computational techniques. Medical image processing tasks include 2D picture visualization, exploration, and 3D volumes, segmentation, classification, registration, and 3D reconstruction of image data [52]. Making visual representations of the body's organs and tissues using digital imaging technology improves a doctor's capacity to make a medical diagnosis. [55].

With the help of technology from the European Medicines Agency, clinicians can treat patients in more remote regions with equipment that is more mobile than an MRI or CT scanner. The cost savings offered by European Medicines Agency technology may make cutting-edge medical imaging technology accessible to a large number of individuals in underdeveloped nations [58]. Enhancement, restoration, encoding, and compression of images are frequently used image processing techniques [62]. Medical image processing includes the use and investigation of human body 3D image collections, typically from a CT or MRI scanner, to identify diseases, direct medical treatments like surgical planning, or for study [65]. Medical image classification's primary goal is to determine which body areas are affected by a disease, not just to achieve high accuracy [67]. The future of medical imaging depends on technological advancements that enhance deeper cell-level visualization and computing speed. As tools for diagnosing and interpreting medical images, AI and ML will become more prominent [69]. X-ray, CT, and MRI are the imaging modalities that are utilized the most commonly. X-ray and CT both need the use of ionizing radiation, whereas MRI detects body protons using a magnetic field. Even though each technology has advantages, MRI is the safest of the three [93].

A variety of biological signals, including EEG, and ECG, are used in signal-based approaches [75]. The following are some benefits of employing digital signals, including digital signal processing and communication systems: Less noise, distortion, and interference can be used to transmit information via digital signals [77]. Digital circuits are easily and inexpensively reproducible in large quantities [78]. The output of a thermocouple transmits temperature information, the output of a pH meter transmits acidity information, and motion, sound, photos, movies, and biological membrane potentials are examples of signals [81]. The advantages of digital signals include higher carrier wave frequencies that transfer more information per second than analog signals, better quality over longer distances than analog signals, automatic operation, ease of noise removal, and potential noise immunity [84].

Biomedical signals that are electric and magnetic come from electromagnetic sources that are located inside the body. To interpret the optical characteristics of the biological system, optical biomedical signals can be measured [85]. Common signal words are those that emphasize, add to, compare to, contrast with, illustrate, and indicate cause and effect [86]. Speech recognition, video streaming, cellular networks, and MRI scans are a few examples of systems that alter signals. The analysis and synthesis of signals as well as their interaction with systems are topics covered by the fields of signal and image processing [88]. Bio-signals, also known as biological signals, are records of a biological event, such as a beating heart or a muscle contracting, in space, time, or space-time. These biological processes frequently result in signals that are monitored and studied due to electrical, chemical, and mechanical activity [89].

4.5 Discussion

Quantitative, image-based, and signal-based approaches are the three most popular methods in healthcare. For example, fMRI is more expensive than EEG and quantitative data, but it produces superior findings. EEG data can also be easily collected, especially for patients. The framework is covered in depth in this thesis, including what types of features are retrieved, which classifiers are best for a particular approach, and which one is more accurate than the others.

The results in table 4.4 indicate the general disparities between three commonly used healthcare techniques. A quantitative-based technique, an image-based method, and a signal-based method are the three different aspects. Datasets in tabular form are acquired from any hospital or social media application in quantitative-based methodologies. X-ray, CT scan, and FMRI datasets were included in the image-based technique, and they were represented in image form. Those authors employed image-based clinical data, and their dataset trials were restricted. The last technique is signal-based, and it includes datasets like ECG and EEG. The comparison of the three current methodologies is innovative in and of itself; the findings of this study suggest the suitable framework within the scientific community.

This study have some insights that the quantitative-based method future predictions are lot of scope discover efficient model, it helps to compare with other studies. Present user-based healthcare data and freely available datasets. The limitations are KNN best for training, slow on large dataset to evaluate. It needs comparison with clinical results is fruitful for implementations. The image-based

method implemented on graphical processing unit decrease computational time. Less time required for test and low cost of X-ray. Tissue abnormality detection is important task using x-ray and this method identifying peoples at high risk to take necessary actions related healthcare. On the other hand limitations of image data are mostly applied on small datasets tested and MRI and CT scan are very costly expenses. The last signal-based method ML method beneficial for larger datasets and less time required for test. Its limitation is failure of the sensors and EEG cost is too high in medical healthcare field.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

Finally, this research covered every related areas of DM in medical healthcare and three distinct approaches in depth. The contribution of this research, its limitations, future work and summary is discussed in the following sections. The overall summary of the research contributions produced by this study is presented in this chapter, along with potential future research areas that could be explored to improve the current research field.

5.2 Conclusions

In this study, SLR was done on studies that discussed the present state of medical treatment and DM procedures. Three other applied approaches that are most commonly employed in the medical area are also described here. This study examines prior literature to assess diabetes in medical healthcare by identifying keywords and research topics. The goal was to verify the current status and novelty that compare it to the previous results using three methods: quantitative, image-based, and signal-based. The SLR protocol was created, and ten percent of chosen studies were piloted to ensure that the methodology was accurate. The studies were chosen from well-known digital libraries; for this purpose, 108 papers were consulted based on certain criteria and evenly distributed to get an understanding of the aforesaid methodologies. The articles chosen span the years from 2015 to 2022, to compare and contrast various medical healthcare approaches in terms of ease, cost, accuracy, and efficiency.

The findings of this study were analyzed using NVivo, the most widely used software for qualitative data analysis. These findings show the general differences between three widely utilised healthcare methods. The three different features include a quantitative-based methodology, an image-based method, and a signal-based method. Datasets are collected in tabular form using quantitative approaches for any hospital or social media platform. The image-based technique used datasets from x-ray, CT, and fMRI scans that were represented as images. The authors used clinical data based on

images, and their dataset trials were constrained. The final method is signal-based and uses ECG and EEG datasets. The comparison of three current approaches is novelty, in and of itself, and this study's conclusions point to the most appropriate framework for the scientific community. By comparing and contrasting, this study demonstrates a better framework for data analysts and data science specialists. Knowledge of appropriate DM procedures and the proposed framework aid the medical sector in a more beneficial way.

5.3 Contributions

The initial contribution of this study is to provide an update on the present state of DM, ML, DS, DL, BDA and related methods in medical healthcare. The novelty is comparing three distinct approaches in the second step. Quantitative-based methods, image-based methods, and signal-based methods are the three approaches to compare and identify the components which are common, and beneficial for community. It also identified different components which are specifically occupied by methods and their uses. These are best with in the method but overall best of three components and features provide enhancement in medical field effectively. This research provides a thorough understanding of diseases mellitus DM in medical research and suggested a complete framework. This study is helpful for different organizations like American Medical Association (AMA), Association for the Healthcare Environment (AHE), Health Care Compliance Association (HCCA), Society of Hospital Medicine (SHM), and World Health Organization (WHO) etc.

5.4 Limitations

Any study project will certainly have limitations, which should be reduced or minimized. This study's objective is to conduct an SLR of the research on the DM in medical healthcare. However, this research study has some potential limitations, such as

- The major drawback is even though each effective approach is reviewed and a thorough framework is suggested, but no method is applied and validated. Because this study just focuses on SLR, which provides a comprehensive view of DM approaches in healthcare applications.

- Second is that, relevant concepts are considered like deep learning, data science, and big data analytics but search some keywords other than these.
- These included studies duration “2015 to 2022” and were published in journals only. By increasing time duration can improve scope of research.
- It only includes material that has been published in English-language journal articles, the literature examined for this research study is not exhaustive. However, it's likely that published material relevant to the study's field is also available in other languages but hasn't been observed because of language barriers.

5.5 Future work

This study will aid the researcher in gaining a thorough understanding of DM's role in healthcare and provide a strong foundation for future implementation. It is possible to do so using the framework described in this study. These novel comparisons are quite valuable for the research community in expanding new possibilities for additional inventions by utilizing the framework offered. This comparison can also be validated by applying other research methodologies like case study or experiments. The researchers can expand this research by identifying more relevant keywords. It can be enhanced to exploring different more effective methods other than quantitative, image and signals-based. This can give potential researchers a precise research path to expand the current effort to address a different, related area of the topic.

5.6 Summary

The important contributions provided in this research study have been covered in this chapter. The interested researchers may also examine the suggested next research directions to enhance the current study.

REFERENCES

[1] I. O. Ogundele, O. L. Popoola, O. O. Oyesola, and K. T. Orija, “A Review on Data Mining in Healthcare,” vol. 7, no. 9, pp. 698–704, 2018.

[2] D. Mining, “A Systematic Review on Healthcare Analytics : Application and Theoretical Perspective of Data Mining,” 2018, doi: 10.3390/healthcare6020054.

[3] N. Caballé-cervigón, J. L. Castillo-sequera, and J. A. Gómez-pulido, “applied sciences Machine Learning Applied to Diagnosis of Human Diseases : A Systematic Review,” pp. 1–27, 2020, doi: 10.3390/app10155135.

[4] S. Khanra, A. Dhir, and A. K. M. N. Islam, “Big data analytics in healthcare : a systematic literature review,” *Enterp. Inf. Syst.*, vol. 14, no. 7, pp. 878–912, 2020, doi: 10.1080/17517575.2020.1812005.

[5] H. K. Bharadwaj, A. Agarwal, V. Chamola, S. Member, and N. R. Lakkaniga, “A Review on the Role of Machine Learning in Enabling IoT Based Healthcare Applications,” pp. 38859–38890, 2021, doi: 10.1109/ACCESS.2021.3059858.

[6] M. Manjiri, M. Mastoli, U. R. Pol, and R. D. Patil, “Machine Learning Classification Algorithms for Predictive Analysis in Healthcare,” pp. 1225–1229, 2019.

[7] A. Patel, S. Gandhi, S. Shetty, and P. B. Tekwani, “Heart Disease Prediction Using Data Mining,” pp. 4–6, 2017.

[8] J. Santos-pereira, L. Gruenwald, and J. Bernardino, “Top data mining tools for the healthcare industry,” *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.06.002.

[9] I. Type and M. A. Edgar, “A systematic literature review of data science , data analytics and machine learning applied to healthcare engineering systems,” 2021.

[10] M. Nabeel, S. Majeed, M. J. Awan, and H. Muslih-ud-din, “Review on Effective Disease Prediction through Data Mining Techniques,” no. October, 2021, doi: 10.15676/ijeei.2021.13.3.13.

[11] A. Site, J. Nurmi, S. Member, E. S. Lohan, and S. Member, “Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis,” *IEEE Access*, vol. 9, pp. 112221–112235, 2021, doi: 10.1109/ACCESS.2021.3103268.

[12] J. Polisena *et al.*, “Case Studies on the Use of Sentiment Analysis to Assess the Effectiveness

and Safety of Health Technologies : A Scoping Review," pp. 66043–66051, 2021, doi: 10.1109/ACCESS.2021.3076356.

[13] F. A. Khan, S. Member, K. Zeb, M. Al-rakhami, and A. Derhab, "Detection and Prediction of Diabetes Using Data Mining : A Comprehensive Review," pp. 43711–43735, 2021, doi: 10.1109/ACCESS.2021.3059343.

[14] N. Kobzeva, V. Terentev, and I. Zolotuhina, "Applied aspects of data mining for decision support at the regional health system Applied aspects of data mining for decision support at the regional health system," 2021, doi: 10.1088/1742-6596/2094/3/032006.

[15] M. L. Kolling *et al.*, "Data Mining in Healthcare : Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development," 2021.

[16] A. Suragala and V. Pynam, "A Comparative Study of Performance Metrics of Data Mining Algorithms on Medical Data," no. February 2021, 2020, doi: 10.1007/978-981-15-7961-5.

[17] S. E. Group, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007.

[18] S. G. Alonso, I. De, T. Díez, S. Hamrioui, M. López-coronado, and M. López-coronado, "SYSTEMS-LEVEL QUALITY IMPROVEMENT A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector," 2017, doi: 10.1007/s10916-017-0832-2.

[19] M. Spiliopoulou and P. Papapetrou, "Guest editorial: Special issue on mining for health," *Data Min. Knowl. Discov.*, vol. 35, no. 4, pp. 1710–1712, 2021, doi: 10.1007/s10618-021-00767-3.

[20] R. Raja, I. Mukherjee, and B. K. Sarkar, "A Systematic Review of Healthcare Big Data," vol. 2020, 2020.

[21] L. T. Majnari and F. Babiˇ, "AI and Big Data in Healthcare : Towards a More Comprehensive Research Framework for Multimorbidity," 2021.

[22] A. Gangal, P. Kumar, S. Kumari, and A. Saini, "Prediction Models for Healthcare using Machine Learning : A Review," 2021.

[23] A. Soni, "Survey on Data mining approach and Feature Scope," vol. 8, no. 5, pp. 440–448, 2021.

[24] R. Safdari, S. Rezayi, S. Saeedi, M. Tanhapour, and M. Gholamzadeh, "Using data mining techniques to fight and control epidemics : A scoping review," *Health Technol. (Berl).*, pp. 759–771, 2021, doi: 10.1007/s12553-021-00553-7.

[25] L. Yang, S. Zheng, X. Xu, Y. Sun, X. Wang, and J. Li, "Medical Data Mining Course Development in Postgraduate Medical Education : Web-Based Survey and Case Study Corresponding Author :," vol. 7, pp. 1–18, 2021, doi: 10.2196/24027.

[26] A. Ramachandran and A. Karuppiah, "A Survey on Recent Advances in Machine Learning

Based Sleep Apnea Detection Systems," pp. 1–19, 2021.

[27] F. Ali, S. El-sappagh, S. M. R. Islam, and A. Ali, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Futur. Gener. Comput. Syst.*, vol. 114, pp. 23–43, 2021, doi: 10.1016/j.future.2020.07.047.

[28] A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," vol. 8, no. Ml, 2020, doi: 10.1109/ACCESS.2020.3001149.

[29] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Abstract Cardiovascular disease is one of the biggest cause for morbidity and mortality among the," *Telemat. Informatics*, 2018, doi: 10.1016/j.tele.2018.11.007.

[30] C. Wu and I. Tzeng, "An Innovative Scoring System for Predicting Major Adverse Cardiac Events in Patients With Chest Pain Based on Machine Learning," vol. 8, 2020.

[31] S. Kaur and R. K. Bawa, "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System," vol. 6, no. 4, pp. 17–34, 2015.

[32] C. S. Hoill and J. Kyungyong, "Development of a medical big-data mining process using topic modeling," 2017, doi: 10.1007/s10586-017-0942-0.

[33] C. H. Lee and H. Yoon, "Medical big data : promise and challenges," vol. 2017, no. 1, pp. 3–11, 2017.

[34] C. Wu, C. Lo, C. Tung, and H. Cheng, "Applying Data Mining Techniques for Predicting Prognosis in Patients with Rheumatoid Arthritis," no. 1, pp. 1–12, 2020.

[35] P. Xia *et al.*, "Data Mining-Based Analysis of Chinese Medicinal Herb Formulae in Chronic Kidney Disease Treatment," vol. 2020, 2020.

[36] S. Joshi, M. K. Nair, and Á. Á. Healthcare, "Prediction of Heart Disease Using Classi fi cation Based Data Mining Techniques," vol. 2, 2015, doi: 10.1007/978-81-322-2208-8.

[37] P. Wu *et al.*, "An Effective Machine Learning Approach for Identifying Non-Severe and Severe Coronavirus Disease 2019 Patients in a Rural Chinese Population : The Wenzhou Retrospective Study," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3067311.

[38] R. Abdulqadir, "Data Mining Classification Techniques for Diabetes Prediction," *ieee*, 2021, doi: 10.48161/Issn.2709-8206.

[39] Y. Liu, Z. Yu, and Y. Yang, "Diabetes Risk Data Mining Method Based on Electronic Medical Record Analysis," vol. 2021, 2021.

[40] M. K. Ahirwar, P. K. Shukla, and R. Singhai, "CBO-IE : A Data Mining Approach for Healthcare IoT Dataset Using Chaotic Biogeography-Based Optimization and Information Entropy," vol. 2021, 2021.

[41] T. Ahmadi-jouybari, S. Najafi-ghobadi, R. Karami-matin, and S. Najafian-ghobadi,

"Investigating factors affecting the interval between a burn and the start of treatment using data mining methods and logistic regression," vol. 5, pp. 1–6, 2021.

[42] T. Iyamu and K. Nunu, "Healthcare data management conceptual framework for service delivery," pp. 3513–3527, 2021.

[43] Z. Parsons and S. Banitaan, "Ecological Informatics Automatic identification of Chagas disease vectors using data mining and deep learning techniques," *Ecol. Inform.*, vol. 62, no. March, p. 101270, 2021, doi: 10.1016/j.ecoinf.2021.101270.

[44] V. J. S. Vimal, M. K. Mi, and Y. Lee, "AI - based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.

[45] M. H. Avizenna, R. A. Widyanto, D. K. Wirawan, T. A. Pratama, and A. Nabila, "Implementation of Apriori Data Mining Algorithm on Medical Device Inventory System," vol. 2, no. 3, pp. 55–63, 2021.

[46] R. Nopour, H. K. Arpanahi, M. Shanbehzadeh, and A. Azizifar, "Performance analysis of data mining," pp. 1–8, 2021, doi: 10.4103/jehp.jehp.

[47] T. Chen, J. Rong, L. Peng, J. Yang, G. Cong, and J. Fang, "Analysis of Social Effects on Employment Promotion Policies for College Graduates Based on Data Mining for Online Use Review in China during the COVID-19 Pandemic," 2021.

[48] A. W. Kempa-liehr *et al.*, "International Journal of Medical Informatics Healthcare pathway discovery and probabilistic machine learning," *Int. J. Med. Inform.*, vol. 137, no. December 2019, p. 104087, 2020, doi: 10.1016/j.ijmedinf.2020.104087.

[49] A. S. Panayides *et al.*, "AI and Medical Imaging Informatics : Current Challenges and Future Directions," 2020, doi: 10.1109/JBHI.2020.2991043.

[50] K. Raza and N. K. Singh, "A Tour of Unsupervised Deep Learning for Medical Image Analysis," pp. 1–29, 2018.

[51] A. Manuscript, "MRI-based attenuation correction for brain PET/MRI based on anatomic signature and machine learning," 2018.

[52] S. Basaia, F. Agosta, L. Wagner, E. Canu, and G. Magnani, "NeuroImage : Clinical Automated classi fi cation of Alzheimer ' s disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage Clin.*, vol. 21, no. October 2018, p. 101645, 2019, doi: 10.1016/j.nicl.2018.101645.

[53] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "NeuroImage Machine learning framework for early MRI-based Alzheimer ' s conversion prediction in MCI subjects," *Neuroimage*, vol. 104, pp. 398–412, 2015, doi: 10.1016/j.neuroimage.2014.10.002.

[54] J. Bullock, C. Cuesta-l, and A. Quera-bofarull, “implementation for medical X-Ray image segmentation suitable for small datasets,” 2019.

[55] F. Uysal, F. Hardalaç, O. Peker, and T. Tolunay, “applied sciences Classification of Shoulder X-ray Images with Deep Learning Ensemble Models,” 2021.

[56] R. Kumar, R. Arora, V. Bansal, and V. J. Sahayasheela, “Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers,” pp. 1–10, 2020.

[57] S. Tang *et al.*, “Data valuation for medical imaging using Shapley value and application to a large - scale chest X - ray dataset,” *Sci. Rep.*, pp. 1–9, 2021, doi: 10.1038/s41598-021-87762-2.

[58] P. Saha, M. S. Sadi, and M. Islam, “Informatics in Medicine Unlocked EMCNet : Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers,” *Informatics Med. Unlocked*, vol. 22, p. 100505, 2021, doi: 10.1016/j.imu.2020.100505.

[59] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on,” *Zeitschrift fßr Medizinische Phys.*, vol. 29, no. 2, pp. 102–127, 2019, doi: 10.1016/j.zemedi.2018.11.002.

[60] M. J. Muckley *et al.*, “Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction,” vol. 40, no. 9, pp. 2306–2317, 2021.

[61] M. J. Willemink, W. A. Koszek, M. S. C. Hardell, M. S. J. Wu, D. L. Rubin, and M. S. M. P, “Preparing Medical Imaging Data for Machine Learning,” no. 21, 2020.

[62] W. Mao, X. Chen, and F. Man, “Imaging Manifestations and Evaluation of Postoperative Complications of Bone and Joint Infections under Deep Learning,” vol. 2021, 2021.

[63] B. M. Z. Hameed *et al.*, “Engineering and clinical use of artificial intelligence (AI) with machine learning and data science advancements : radiology leading the way for future,” pp. 1–12, 2021, doi: 10.1177/17562872211044880.

[64] A. Z. Khuzani, M. Heidari, and S. A. Shariati, “COVID-Classifier : An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images Abstract : Introduction : Results :,” no. Ml, 2020.

[65] J. Rasheed, A. Ali, H. Chawki, D. Akhtar, J. Fadi, and A. Turjman, “ORIGINAL RESEARCH ARTICLE A machine learning - based framework for diagnosis of COVID - 19 from chest X - ray images,” *Interdiscip. Sci. Comput. Life Sci.*, vol. 13, no. 1, pp. 103–117, 2021, doi: 10.1007/s12539-020-00403-6.

[66] M. A. Elaziz, K. M. H. Id, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, “New machine

learning method for image- based diagnosis of COVID-19," 2020, doi: 10.1371/journal.pone.0235187.

[67] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID - 19) using X - ray images and deep convolutional neural networks," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1207–1220, 2021, doi: 10.1007/s10044-021-00984-y.

[68] P. Afshar *et al.*, "COVID-CT-MD : COVID-19 Computed Tomography (CT) Scan Dataset Applicable in Machine Learning and Deep Learning," pp. 1–9, 2020.

[69] Y. O. O. J. Ha, G. S. Member, S. Yoo, and S. Member, "Spatio-Temporal Split Learning for Privacy-Preserving Medical Platforms : Case Studies With COVID-19 CT , X-Ray , and Cholesterol Data," *IEEE Access*, vol. 9, pp. 121046–121059, 2021, doi: 10.1109/ACCESS.2021.3108455.

[70] S. Hijazi and A. Page, "Cardiac Health Monitoring and Decision Support," 2016.

[71] T. Tuncer, S. Dogan, P. Pławiak, and U. R. Acharya, "Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals," *Knowledge-Based Syst.*, no. December, p. 104923, 2019, doi: 10.1016/j.knosys.2019.104923.

[72] V. Ponciano, I. M. Pires, F. R. Ribeiro, N. M. Garcia, E. Zdravevski, and P. Lameski, "Machine Learning Techniques with ECG and EEG Data : An Exploratory Study," 2020.

[73] N. Surantha, T. F. Lesmana, and S. M. Isa, "Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data," *J. Big Data*, 2021, doi: 10.1186/s40537-020-00406-6.

[74] V. Ponciano, I. M. Pires, F. R. Ribeiro, M. C. Teixeira, and E. Zdravevski, "Experimental Study for Determining the Parameters Required for Detecting ECG and EEG Related Diseases during the Timed-Up and Go Test," 2020.

[75] W. He and G. Wang, "Simultaneous Human Health Monitoring and Time-Frequency Sparse Representation Using EEG and ECG Signals," vol. 7, pp. 85985–85994, 2019.

[76] M. Wasimuddin, K. Elleithy, and S. Member, "Stages-Based ECG Signal Analysis From Traditional Signal Processing to Machine Learning Approaches : A Survey," vol. 8, 2020, doi: 10.1109/ACCESS.2020.3026968.

[77] M. A. García-ramírez, "Machine Learning for personalised stress detection : Inter-individual variability of EEG-ECG markers for acute-stress response Computer Methods and Programs in Biomedicine Machine Learning for personalised stress detection : Inter-individual variability o," *Comput. Methods Programs Biomed.*, vol. 209, no. October, p. 106314, 2021, doi: 10.1016/j.cmpb.2021.106314.

[78] R. E. C. G. Classification and F. Interoperability, "Personalized Wearable Systems for Real-

time ECG Classification and Healthcare Interoperability," pp. 9–14, 2017.

[79] F. P. George, I. M. Shaikat, and P. S. Ferdawoos, "Recognition of emotional states using EEG signals based on time-frequency analysis and SVM classifier," vol. 9, no. 2, pp. 1012–1020, 2019, doi: 10.11591/ijece.v9i2.pp1012-1020.

[80] J. Breitenbach and R. Buettner, "Detection of Excessive Daytime Sleepiness in Resting-State EEG Recordings : A Novel Machine Learning Approach Using Specific EEG Sub-Bands and Channels," no. August 2021, pp. 0–9, 2020.

[81] I. Ullah, M. Hussain, E. Qazi, and H. Aboalsamh, "An automated system for epilepsy detection using EEG brain signals based on deep learning approach," *Expert Syst. Appl.*, vol. 107, pp. 61–71, 2018, doi: 10.1016/j.eswa.2018.04.021.

[82] M. Sharma, J. Tiwari, V. Patel, and U. R. Acharya, "Automated Identification of Sleep Disorder Types Using Triplet Half-Band Filter and Ensemble Machine Learning Techniques with EEG Signals," 2021.

[83] S. Raghu and N. Sriraam, "Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms," *Expert Syst. Appl.*, 2018, doi: 10.1016/j.eswa.2018.06.031.

[84] L. Chen, Y. Zhao, J. Zhang, and J. Zou, "Expert Systems with Applications Automatic detection of alertness / drowsiness from physiological signals using wavelet-based nonlinear features and machine learning," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7344–7355, 2015, doi: 10.1016/j.eswa.2015.05.028.

[85] O. Alshorman, M. Masadeh, B. Bin Heyat, F. Akhtar, and H. Almahasneh, "Frontal Lobe Real-time EEG Analysis using Machine Learning Techniques for Frontal lobe real-time EEG analysis using machine learning techniques for mental stress detection Neuroscience Integrative," no. August, 2021.

[86] C. Venkatesan, P. Karthigaikumar, and A. Paul, "ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications," *IEEE Access*, vol. 6, pp. 9767–9773, 2018, doi: 10.1109/ACCESS.2018.2794346.

[87] U. Kumar and S. Yadav, *Application of Machine Learning to Analyse Biomedical Signals for Medical Diagnosis*, no. June. 2021.

[88] Y. Choi *et al.*, "applied sciences Machine-Learning-Based Elderly Stroke Monitoring System Using Electroencephalography Vital Signals," 2021.

[89] D. E. Centre, P. J. Karoly, E. Nurse, and M. J. Cook, "Machine learning and wearable devices of the future," no. July, 2020, doi: 10.1111/epi.16555.

[90] S. K. Pandey and R. R. Janghel, "Recent Deep Learning Techniques , Challenges and Its

Applications for Medical Healthcare System : A Review," *Neural Process. Lett.*, no. Ml, 2019, doi: 10.1007/s11063-018-09976-2.

[91] S. Purushotham, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *J. Biomed. Inform.*, vol. 83, no. June, pp. 112–134, 2018, doi: 10.1016/j.jbi.2018.04.007.

[92] X. Zeng, S. Lin, and C. Liu, "Multi-View Deep Learning Framework for Predicting Patient Expenditure in Healthcare," vol. 2, no. December 2020, 2021, doi: 10.1109/OJCS.2021.3052518.

[93] X. Images, "COVID-19 Detection Using Deep Learning Algorithm on Chest X-ray Images," 2021.

[94] N. Musa, A. Ya, N. Aljojo, H. Chiroma, and K. S. Adewole, "A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram," 2022.

[95] J. Robert, Y. Zhang, and J. Gwizdka, "International Journal of Medical Informatics Healthcare professionals ' acts of correcting health misinformation on social media," vol. 148, no. November 2020, 2021.

[96] S. Nazir *et al.*, "A Comprehensive Analysis of Healthcare Big Data Management , Analytics and Scientific Programming," vol. 8, 2020, doi: 10.1109/ACCESS.2020.2995572.

[97] A. A. Nancy, D. Ravindran, P. M. D. R. Vincent, K. Srinivasan, and D. G. Reina, "IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning," 2022.

[98] M. M. Rathore, A. Ahmad, A. Paul, J. Wan, and D. Zhang, "Real-time Medical Emergency Response System : Exploiting IoT and Big Data for Public Health," *J. Med. Syst.*, 2016, doi: 10.1007/s10916-016-0647-6.

[99] P. Sarosh, S. A. Parah, G. M. Bhat, and K. Muhammad, "A Security Management Framework for Big Data in Smart Healthcare," *Big Data Res.*, vol. 25, p. 100225, 2021, doi: 10.1016/j.bdr.2021.100225.

[100] M. Viceconti, P. Hunter, and R. Hose, "Big Data , Big Knowledge : Big Data for Personalized Healthcare," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 4, pp. 1209–1215, 2015, doi: 10.1109/JBHI.2015.2406883.

[101] D. Piovani and S. Bonovas, "Real World — Big Data Analytics in Healthcare," pp. 10–12, 2022.

[102] R. Salazar-reyna, F. Gonzalez-aleu, E. M. A. Granda-gutierrez, J. Diaz-ramirez, and J. A. Garza-reyes, "A systematic literature review of data science , data analytics and machine learning applied to healthcare engineering systems systems," 2020, doi: 10.1108/MD-01-2020-

0035.

[103] A. Garg and V. Mago, “Role of machine learning in medical research : A survey,” *Comput. Sci. Rev.*, vol. 40, p. 100370, 2021, doi: 10.1016/j.cosrev.2021.100370.

[104] A. Dhillon and A. Singh, “Machine Learning in Healthcare Data Analysis : A Survey,” vol. 8, no. June, pp. 1–10, 2019, doi: 10.15412/J.JBTW.01070206.

[105] R. Buettner and N. Bay, “A Systematic Literature Review of Medical Image Analysis Using Deep Learning,” pp. 40–43.

[106] A. Pashazadeh and N. J. Navimipour, “Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review,” *J. Biomed. Inform.*, 2018, doi: 10.1016/j.jbi.2018.03.014.

[107] M. M. Malik, S. Abdallah, and M. Ala, “Data mining and predictive analytics applications for the delivery of healthcare services : a systematic literature,” *Ann. Oper. Res.*, 2016, doi: 10.1007/s10479-016-2393-z.

[108] S. Imran, T. Mahmood, A. Morshed, and T. Sellis, “Big Data Analytics in Healthcare — A Systematic Literature Review and Roadmap for Practical Implementation,” vol. 8, no. 1, pp. 1–22, 2021.