EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 154 No. 11
November 2025

Research in Computing Science

Research in Computing Science, Año 24, Volumen 154, No. 11, noviembre de 2025, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de noviembre de 2025.

Research in Computing Science, year 24, Volume 154, No. 11, November 2025, is published monthly by the Center for Computing Research of IPN.

Volume 154(11)

Advances in Artificial Intelligence

Grigori Sidorov (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2025

ISSN: in process

Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Data Collection and Analysis Techniques in Computer Science Research: A Review

Muhammad Ateeb Ather

Department of Computer Sciences, Bahria University, Lahore, Pakistan

03-134211-022@student.bahria.edu.pk

Abstract. Many researchers and scientists face difficulties in accomplishing research work successfully due to the lacking of good knowledge regarding method of scientific data collection. Several methods are utilized for collection of data in research work. Effectiveness of every research depend on the validity of collected data. Appropriate research questionnaires provide motivation and direction in structuringdata gathering and data analysis approaches. This paper focuses on the data collection methods and data analysis techniques in terms of qualitative and quantitative particularly for computer scientists. Furthermore, this work is to examine numerous analysis methods, techniques and tools that could beapplied to diverse type of data science applications.

Keywords: Data collection, data analysis, qualitative research, quantitative research.

1 Introduction

This research is basically a dedicated activity which is more than gathering information or report writing. It contains information gathering in a traditional systematic way, which can be examined methodically to bring to answers of research questionnaires and calculate results. Datacollection is considered to be the core of any research design regardless of the area of study [69].Every research starts with definitive questions, which must be responded. Central rule regarding data collection and analysis is not to gather and examine all types of data. Inappropriate data collection effects on the result of a study and bring towards inaccurate or wrong results. Basically data collection is considered to be the process of collecting and evaluating information in a targeted systematic manner that provides the facility to someone to respond according to research queries,hypothesis testing and results calculations.

Data gathering research module is common to all study areas including business, social and physical sciences and humanities, etc. Although methods differ by research field, main focus on guaranteeing appropriate and accurate collection must be same. The primary goal behind data collection is to maintain research integrity. Data

collection is said to be challenging task needs throughout the planning, development and analyzing process to complete job successfully. Data collection begins with defining what type of data needed (qualitative and quantitative). Which source of data required (primary and secondary).

By using different techniques data can be gathered. Decision regarding which instrument will be utilize for data collection is conducted via research question. After the gathering of appropriate and accurate data through the usage of suitable technique, next phase is the discovery of useful and appropriate information hidden inside data for further interpretation and manipulation. Analysis word defines operations, using with the intent thatsummarization of gathered data and forming to such an extent yielding responses to queries. In other words, it indicates, study of data to define inherent statistics. Process of applying statistical and logical methods for the extraction of accurate and related information from data is said to be data analysis [39].

During data analysis process, three things play an important role. First is organizationof data, second is the classification and summarization of data, both used for the reductionof data and third is the identification and linkage of data patterns. Researchers do data analysis in bottom up and top-down mode. Data analysis includes forming of data in an appropriate manner.

Problem behind analysis of data differs from area of study to study. Different analysis techniques presented by researchers according to data type. First section describes data and information, data sources and data types. Second section describes different ways of data collection. Third section presents definition of data analysis, methods of data analysis, tools of data analysis and research techniques. Forth focuses on the data analysis techniques in terms of qualitative and quantitative perspectives. Fifth section compares quantitativeand qualitative methods.

2 Data and Information

Data and information both are same notions, but both are not similar thing. Primary difference among both is that information is entire and data is the portion. Information is basically the outcome of processed data. Information word drives from Latin word meaning ‘conception or formation’. Information concept has diverse meanings in diverse type of contexts. Hence concept of information becomes linked to ideas of communication, education, understanding, perception, control and knowledge etc. Information related with data. Information solve out uncertainty.

Information is transferable in time,with the help of space, telecommunication, data storage and communication. Though data and information terms used interchangeably, both terms different meanings. Researchers defined concept of information in different perspectives: information as data in environ, information as resource, information as communication process part and information as knowledge representation [48].

Data word is the plural form of “Datum”. In numerous means, data defined according todiverse types of people according to their specialized area. According to this study, data is stated as gathering of facts, like measurements, observation or values and statements about things[12]. Data word represents information, whichgathered in some organized way and help out someone to understand the information properly. Representation of instructions, concepts and facts in formal way appropriate for

Table 1. Comparison among primary and secondary data.

Features	Primary Data	Secondary Data
Data	Real time data	Past data
Process of collection	Very involved	Easy and fast
Cost of collection	High	Low
Time consumed	More time	Less time
Type of data	Qualitative	Quantitative
Reliability	More reliable	Less reliable
Accuracy	Accurate	Lack of accuracy
Sources	Surveys, observations, experiments	Internal records,Govt, published data etc

processing, interpretation and communication via automatic means and humans is said to be data. Data not gathered irregularly, it is necessary with regards to answer queries [3].

As discussed earlier, data collection is considered necessary in every research area. After the defining of research problem, data collection task starts. Given below section elaborate different types of data and givesclear depiction of sources through which data canbe collected easily.

2.1 Data Sources Types

Data sources can be subdivided into two categories: secondary and primary sources of data.

2.1.1 Primary Data

Gathering of data from information source andfirst-hand experience is said to be primary data. Insimple words, data which is gathered for first timeby scholars/authors himself. It is more realistic, accurate and objective. Its viability is higher than secondary data because it cannot be transformed via humans. Its collection time is longer and always specific to scholar requirements [9].

2.1.2 Primary Data Sources

In primary data collection sources, researcher need to be clearly 1 define population under study also analysis units. Primary data sources are inadequate and due to deficiency of collaborationit's hard to acquire data from primary architect. Primary data mostly gathered via negotiations or face-to-face discussions and through exchanging of emails, radio based communication, phone oriented discussions and direct surveillance. Otherprimary data architects are observations, survey, interviews, experiments and questionnaires [40].

2.1.3 Secondary Data

A way in which data gathered from architect which has been already advertised is known as secondary data. Literature review found in each research is considered to be

the foundation ofsecondary data. For social sciences, secondarydata architects involve organizational reports and gathering of data from qualitative based research and quantitative approaches. It is said to be less effective. In some cases, when difficult to captureprimary data, attaining information from secondary architect is feasible. Usage ofsecondary data is less costly and quicker in terms of distinguishing to primary data [9].

2.1.4 Secondary Data Sources

There are few means of gathering secondary data involve newspaper, books, research and internet articles, records, databases and profiles etc [40].

2.2 Data Types

There are two types of data: qualitative and quantitative.

2.2.1 Qualitative Data

This type of data is not in the form of numbers,basically descriptive. Most of the times, such dataexpresses sentiments, moods and feelings etc. It issaid to be the depiction of conditions, actions, connections, direct citations, casestudies andrecords. Qualitative studies address questions ‘what’ ‘why’ and ‘how’. This type of data gathering procedures play necessary part in influence estimation via supplying information beneficial to comprehend methods beyondobserved outcomes and evaluate variations in people sensation of their welfare. Furthermore, it is costly and takes time. Scholars need to capture data systematically and accurately. Data gatheringapproaches need to perceive ethical values of research. Qualitative approaches mostly utilized inevaluation and can be organized in three groups: methods of observation, in-depth interview and document review [83].

2.2.2 Quantitative Data

Type of data which can be expressed in terms of numbers. By nature, it is in the form of structured and unstructured. Arranged or organized form of data is said to be structured data,while un-organized form of data called unstructured data. Via closed queries structured data can be generated, unstructured data can be generated via open queries [12,13]. It utilizes various measurement scales like ordinal, nominal,ratio and interval scale. Quantitative studies address questions ‘how many’ and ‘how much’. Quantitative data collection approaches generate outcomes which can be easily summarized, compared and generalized.

2.2.3 Mixed Methods

This method gathers qualitative and quantitative data within similar study. Several researchers assume that mixed method is a new technique but researchers have been gathered qualitative and quantitative data for several eras. According to [37] when studying difficult andcomplex issues, one method just show a slight vision of entire depiction. Mixed methods deliverinfo on various stages of understanding.

Qualitative methods give deep understand ability of variables which takes to quantitative mathematical findings, when both methods grouped. Mixed methods uses both methods qualitative and quantitative either sequentially or simultaneously [79]. It

increases research validityand reliability. Researchers face problems when they use mix methods technique is that, it is time consuming, expensive and takes effort.

2.3 Types of Quantitative Data

Quantitative data classified into two types: discrete and continuous data.

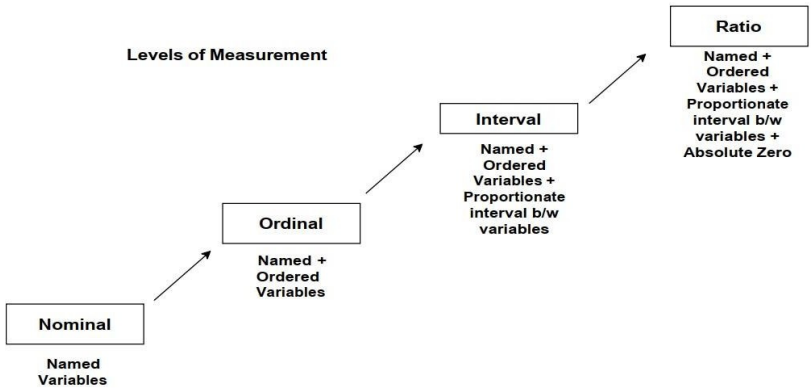

Fig. 1. Quantitative Levels of Measurement: Ordinal, Interval, Nominal and Ratio.

2.3.1 Discrete Data

Type of data which comprises of calculating numbers only. It only uses separate and distinct values. It holds only finite numbers, sub-section ofthose numbers cannot be happened. It holds only those type of values which are in countable form, means impossible to breaking values. Graphical representation of discrete data can be done by bargraph. Its nature is countable. Examples include: no of employs in a company, no of animals in a zoo and no of cars inside parking area etc.

2.3.2 Continuous Data

Type of data which is measurable. It can be reduced in sub-portions. This type of data cascades within continuous order. Graphical representation of continuous data can be done by histogram. Its nature is measurable. For various types of hypothesis tests, this data is use able.

Examples are weight, height, temperature andlength [58].

2.4 Measurement Levels

Measurement can be grouped into four stages:ordinal, interval, ratio and nominal.

2.4.1 Nominal Level

For identification goals, this scale is used. Another name of this scale is “categorical scale” because for easy identification, it allocates quantities to attributes. Quantities behave like labels. It can be examined via pie and bar charts. Example: sex (male/female) and status (single/married) are two groups with categories, but these groups having no ranking or naturalordering [18].

2.4.2 Interval Level

It contains principle of static space among 4 and 5 and 5 and 6, must be equal space. Calendar time is an example of interval scale. It means timeamong 3:00 am and 4:00 am is equal and same tothe 4:00 am and 5:00 am time [22]. Most of the attitude scales are viewed an interval like multi- graphical scales, intelligence scores and semantic differential scales. It is applicable in numerous disciplines such as medicine, education andengineering etc.

2.4.3 Ordinal Level

It contains organizing or ranking of characteristics reliant on the variable being scaled. Articles within this scale ordered according to theamount of existence of variable in query. Characteristics or attributes on this scale generallyorganized in descending or ascending way. It calculates the ratio of existence of variable. This scale applied in advertising, surveys related toclient satisfaction and market based research. For the representation of a degree, it utilizes modifier such as highly, less, very and more etc. Example: ordinal levels mostly found in questionnaires for quality ratings (good, poor, excellent and very good etc.) and agreement (agree, disagree andstrongly disagree etc.) etc [18,87].

2.4.4 Ratio Level

It is said to be the peak level of measurement and it comprises all attributes of zero origin, fixedspace, order and categorization. This scale of measurement permits scholar to match both the variances and comparative magnitude of quantities. Money is an example of ratio level [22,87].

3 Data Collection Methods

By using several methods and from various sources, data can be gathered and attained. Following are the methods used for data collection.

3.1 Observation

It is said to be the most useable method particularly in behavioral sciences studies. For researcher's, it has been become a scientific instrument. Observation is referred to as the "logical depiction of conducts, events and objects in social setting selected for study" [50]. Scholar perceives participants conduct and record outcomes of these perceptions. This method helps out researcher to capture non-verbal signs like emotions, way of communication and note down time they devote in various tasks [60].Furthermore,it is expensive because its time consuming and observer needs to be stay during the whole process.

3.2 Focus Groups

This method belongs to qualitative method.Marketing professional Ernest Ditcher and psychologist (1907-1991) invented the term. In this method, group of people are invited

to take part in a conversation on a specific topic. By usingthis method, data type which can be attained contain views, beliefs, and agreement and disagreement expressions with other members and procedures in which single or group personalities are constructed [59]. It provides an open atmosphere, with the intent that members feel free to reveal their opinions and views. To find out results, gathered data examined systematically andthoroughly. Problem with this method is that, when group mediator takes part in discussion maydisturb the flow of discussion among group participants [31]. Initially, for market based research, this method used, but now it is applied in political analysis, heath based campaigns and public sector marketing etc.

3.3 Interview

Verbal method of asking questions is said to beinterview. It's a sequence of queries a scholar asksto someone. Primarily, utilized to acquire an understanding of aims and causes for people conduct, likings, attitudes, views and beliefs on related topics. Interviews can be conductedindividually or in a group [26].

Interviews can be in the form of semi-structured, structured or unstructured. In order to attain data, structured interview depends on the set of planned and standardized questionnaires while unstructured interview does not depend on the set of standardized questions. In semi-structured interview, interviewer utilizes an interview leader and it is basically a set of queries which must be covered during conversation or communication.

Furthermore, it provides free environment for asking queries in any sequence [81]. Interview method is further sub-divided into two categories: face to face and telephone interviews [13].

3.3.1 Face-to-Face Interviews

Also known as in-person interview. It is the most common and oldest way of collecting data. In this method, facial expressions and feelings canbe clearly noticed. This method enhances responserate and guarantees quality of attained data. Though face-to-face interviews are effective but itis costly and takes time [63].

3.3.2 Telephone Interview

Interviewer communicates on telephone in accordance with pre-arranged queries. For thistype of interviewing, closed-ended queries mostly suggested. Basically, it is short and concentratedon a gathering of focused information [64].

3.4 Case Study

For researchers, case study is usually the mostcommon and useful method [5]. It is a deep investigation process about person, condition andgroup etc. It contains gathering of data from different sources. Whenever deep understandability of problems required, case studymethod conducted. Mainly, it is applied in social science studies and address issues relevant to sociology, education and society related problemslike unemployment, illiteracy and poverty etc. It istime consuming and cannot be replicated [85].

3.5 Survey

Survey method is most commonly used to examine feelings, beliefs and opinions. It comprises of discovering facts in specific area of inquiry. Surveys can be directed in different ways: via telephone, face-to-face, through mail and electronically. For gathering data, surveys appliedin several fields even in private and public zones. Researchers can conduct this method in different means based on the selected approach and study purpose. Survey method is time consuming.

3.6 Questionnaire

Most common technique of gathering data from large group of persons. It comprises of questionnaires set hand over to communicator forsolutions. It can be open-ended and closed-ended.Closed based questionnaires provide finite set of answers. In simple words, it contain some sort of scale like ordinal, nominal, ratio and interval and multiple choice type questions. In an openended questionnaire, data gathered can be concentrated on group methodologies furthermore group attributes. In which questions never based upon tick mark but there is blank space where respondent write answer. It is helpful to simplify and compute people conducts and attitudes [12,13].

3.7 Oral History

Process of recording and storing data which is gathered from first hand person from their past memories and experiences. Interviewer asked questions and respondent responses are documented, which stored for future usage.Researcher uses audio-video tapes for interviews.The main difference among interview and oral history is that, interviews concentrated on a specific topic while oral history based on particular topic and it is less organized. Knowledge gained by this method is unique, because it shares the interviewe thoughts, beliefs,views and understanding.

3.8 Document Review

It is a process of gathering data via reviewing current documents. Documents could be external or internal to a firm or a program. Documents canbe in hard form and may be consists of newsletter,program logs, meeting time and reports etc. it is abest source of gathering historical data. Relatively,it is less costly. It takes time to gather, examine and review several documents. Also, data gatheredvia this method can be inappropriate or incomplete[74].

3.9 Ethnography

It's a way of study of people and cultures in a systematic manner. In social sciences discipline, ethnography assumed to be the main method of data collecting. Basically provide deep understanding of rituals, behaviors and likings of different cultures and societies. The main goal of this method is to attain complete understandability of people's beliefs and activities, as well as the location type in which they reside, by

means of perceiving and questionnaires. Researcher directly contributes in data collection process [8].

3.10Experiment

A way of data collection which provide the understandability of cause and effect associations of variables. In other words, affected variable is said to be dependent variable while the variable which will manipulate the dependent variable is known as independent variable. This method applied in various research domains like sociology, agriculture and medical etc. Problem of using experiment method is that it is costly and takes time.

4 Data Analysis

Analysis term denote processes through whichdata organized and summarized in such a way, that leads answer to research questionnaires. For further manipulations, data analysis organizes datain a suitable manner. Data analysis problem differsaccording to research study. Data analysis play a vital role in research domains and it is the process of placing statistics and facts to resolve research issues. Moreover, necessary phase of research is the exploration of data, which is getting from dataanalysis and helpful to create implications and extract outcomes [77]. Data analysis is a method to get order, structure and provide meaning to pileof gathered data. Although this method takes time, disorganized and ambiguous but also calledcaptivating and innovative method [51].

Data analysis is the way of cleansing, reviewing, altering and forming data along with the aim of finding valuable information. Itsupports businesses to make decisions. In the present era, data analysis applied in various domains like science, business and education etc. Also, provide means of effectively performingbusinesses by proper decision making. To ensure integrity of data is the main component of data analysis.

4.1 Data Analysis Types

There are several types of analysis provided along with diverse kind of goals.

4.1.1 Descriptive Analysis

The main goal of this analysis is to summarizeand define data set. For conducting statistical analysis, Descriptive analysis considered to be themain step. It provide means of data distribution, help out to identify outliers and errors and generate correlations among variables. Best way to conduct this analysis is that, first need to decideabout variable types after that select approachwhich will suitable according to types of variable.If descriptive analysis structured systematically, then it takes no time and easy to use [75].

4.1.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is a critical procedure of conducting initial examinations on data in order to identify data patterns, discover variances and hypothesis testing through the assistance of graphical illustrations [42]. The goalbehind

EDA is to discover new aspects of data. Also helpful to explore data and formulate hypothesis which further lead to experiments and collection of new data. For data visualization, some graphical methods are histogram, box and scatter plot etc.

4.1.3 Inferential Analysis

From data samples, inferential analysis permits to make inferences about large population. It comprises of two parts: hypothesis testing and estimating parameters. Hypothesis testing tells about where researcher or someone make usage ofsample data to answer research questionnaires. Inestimating parameters part, it takes statistics from data samples and utilize it to answer about parameters of population.

4.1.4 Predictive Analysis

For making future predictions, predictive analysis uses background or present data to discover patterns. Main objective is the usage of machine learning approaches and statistical algo's to assess future results which depends on past data.Predictions accuracy focused on the input variables. Organizations and companies used predictive analysis to enhance operations and productivity skills [46].

4.1.5 Confirmatory Data Analysis

Confirmatory data analysis (CDA) emphasizes on approving and negating current hypothesis. CDA, in which research examine evidence by utilizing conventional statistical instruments like dependence, importance and inference. It contains many things such as regression and variance analysis, hypothesis testing and to generate approximations along with quantified scale of precision. It is used to fail or approve theory of measurement.

4.1.6 Casual Analysis

Based on the cause and effect associations between variables. Cause denotes a reason. Goal of casual analysis is the identification of the origin of a problem rather than discovering symptoms. Itis helpful to reveal facts which takes toward certain condition or state. With the support of different methods, casual analysis can be formulated.

4.1.7 Mechanistic Analysis

The goal of MA is the understandability of exact variations in variables that bring towards to the variations in other variables. Useful inengineering and physical sciences, needed high fidelity Measurement error is said to be the noise in data. Mechanistic analysis demands maximum effort. Applications which comes under this analysis type are randomized trial set of data.

4.2 Data Analysis Methods

There are different analysis methods accordingto business and technology perspectives, but two main methods are: qualitative and quantitative.

4.2.1 Qualitative Analysis

Qualitative analysis method focused on thequestions like 'how', 'what' and 'why'. Such type of queries are addressable by qualitative means like attitude rating and

questionnaires etc. Qualitative analysis mostly conducted in texts and descriptions means, also contain audio and video illustrations.

4.2.2 Quantitative Analysis

Quantitative analysis method utilizes statistical and mathematical modeling. It is measureable in number forms. Quantitative analysis make usage of statistics to find out answers of research queries in terms of numbers and define the purpose of numbers in respond to research related questions. Some other methods are given below.

4.2.3 Text Analysis

Text analysis is a method of data analysis to gain machine understandable facts and to identifydata patterns from large sets of data by the usage of data mining techniques and via databases. The goal behind this method is the generation of structured data from unstructured content. This method is said to be information extraction, text mining and text analytics. Vagueness of human dialect is considered to be the major challenge of text analysis. There are variety of text analysis software tools. Some of them are: apache OpenNLP, visual text, genism, distributed machine learning toolkit, google cloud natural language API, KNIME text processing and general architecture for text engineering-GATE etc.

4.2.4 Statistical Analysis

Statistics comes under the mathematics branch and deal with data gathering, data interpretation, data analysis and data demonstration whereas biostatistics belongs to statistics branch where statistical approaches applied on biomedical data that lead towards ultimate solution. Statistical analysis play a vital role and considered necessary for those types of researches which contains statistics as a part of research methodology. It demonstrates 'what happen?' via usage historical data. It contains data gathering, data analyzing, data interpreting and presenting of data. Statisticalanalysis examines data set or data sample. It is applicable in situations such as collection of research expositions, plotting studies and surveysand statistical modeling [72]. It is suitable for BI related organizations because these organizations work with huge volumes of data. Statistical Analysis can be classified into two types: descriptive and inferential analysis.

Inferential Analysis

In inferential analysis, analysts examines sample from whole data. Via choosing distinct samples, analysts identify diverse type of conclusions from similar type of data. Also, it shows associations among various variables and make estimations.

Descriptive Analysis

In descriptive analysis, analysts examines whole data. For continuous type of data, it demonstrates deviation and mean while for categorical data type, illustrates frequency andpercentage.

4.2.5 Diagnostic Analysis

It demonstrates 'why did it happen?'. Diagnostic analysis give in-depth analysis to answer queries. To discover behavior patterns of data, diagnostic analysis is feasible. It

is also known as root cause analysis and it contain procedures such as data mining, data discovery, dill down and correlations. In discovery procedure,analysts discover sources of data, then those data sources help out to understand outcomes. Drill down procedure concentrates on the certain data aspect. Data mining procedure basically extract useful information from huge size of data sets. Correlation process denotes a degree in which variable pairs connected linearly [38].

4.2.6 Predictive Analysis

To identify future trends, PA uses historical data and give that data to machine leaning model.To predict what will happen in the future, model applied to existing data. Several companies and organizations choose predictive analysis to attain benefits. It contain methods like data mining, artificial intelligence, data modeling, machine and deep learning. Common uses include fraud detection, risk reduction and operations refinement etc. [38, 46].

4.2.7 Prescriptive Analysis

Prescriptive analysis is said to be a mixture of business rules and data. Data used for prescriptive analysis can be in external and internal forms while business rules are boundaries, best practices and preferences. In this analysis, analysts define which necessary action need to take to resolve present issue. Many organizations prefer prescriptive analysis because it expand data performance. Analysts examine data and take necessary decisions to look at current issues and situations [38].

4.3 Data Analysis Tools

Data analysis tools help users in processing, analyzing and manipulation of data, also examineassociations between sets of data. Furthermore, helpful to discover trends and data patterns for making interpretations and draw conclusions. Selection of right data analysis tool is a challenging task because usage of same tools never fulfills necessary requirements in different domains. First of all, need to understand what typeof data will be used for analysis purpose? After that, must be clearly define, in which form data will be presented because some studies concentrate on qualitative data and some work with quantitative data. So, different analysis toolsare available for diverse type of data. Few of themare given below.

4.3.1 Quirkos

It helps users in sorting, managing and understanding text data. Users utilize Quirkos to code or tag related data units and match themes over hundreds of diverse sources. Furthermore, Quirkos supports qualitative researchers in analyzing and understanding data like articles,surveys and interview scripts. Some features of this program are: visual interface shapes coding easy and fast; execute comparative interpretationsthat highlight variances among respondent clusters and construct multifaceted queries to drilldown within data; coding with Quirkos is simple and directly import pdf, word files and text. Moreover, Quirkos has built in functionality of working with several languages and every script type. Quirkos is a good mode of work with small data. In Quirkos, everything is attractive, interesting, visual and colorful.

4.3.2 Nvivo

For qualitative and mixed method research,Nvivo provide support. It's a qualitative data analysis software. Nvivo is helpful for organizing, analyzing and discovering insights in qualitative and unstructured data like open-ended surveys, social media, interviews, web content and articles.Without Nvivo, when work with qualitative data, work will take time and difficult to navigate. Somefeatures are: examine and import video, web data, images, online surveys and emails; coding, word frequency, text search, matrix coding and coding comparison queries; coding review; relationship coding; code and organize several sources of datainto one main file; add notes and interpretations; data visualization and assign attributes to data. User interface and text analysis presented in different languages such as French, Spanish, English, Chinese, Japanese and English etc. For windows and MAC OS, Nvivo is suitable. It savestime and speedily retrieve, store and organize data[6]. Researcher presented five necessary tasks through which Nvivo simplify qualitative dataanalysis. Tasks are: query and manage data, reporting, manage ideas and visual based modeling [33].

4.3.3 WebQDA

In distributed and collaborative environ, WebQDA software provide support for qualitative data analysis. Though in qualitative analysis, there are some software packages which address unstructured (image, audio etc.) and nonnumeric data and some others packages can be utilized various researchers in a distributed and collaborative work environment.

WebQDA software directed researchers in diverse situations that require analysis of qualitative data, asynchronously and synchronously, collaboratively and separately. Features ofwebQDA are: simple in usage; provide security ofdata; feasible for several research kinds; web-based software and well-suited for all OS's. It follows other programs design that accessible in market and creates difference via providing real-time, online cooperative work and service which assists process of research. With the help of WebQDA, documents editing, linking, viewing and organizing can easily be done [17].

4.3.4 HyperRESEARCH

For qualitative analysis, it is considered to be the powerful tool. It is reliable and flexible. Tool is simple to learn and use, cost-effective and cross-platform. Without usage of additional powerful features, researcher can easily learn how to retrieve data and data coding. Cross-platform abilities allow consumers to utilize it anyplace or on every system (computer). It provides innovative features which give control and whole accessibility of data. Also, provides flexibility in respect of how someone can get access of data. HyperRESEARCH caseoriented architectureregives adaptability to auto-code several sources to several cases, applying prescribed codes to different phrases and keywords via single command. Code map construct visual illustration of codes and their associations and behave just asa robust code selection tool. With the help of statistical software, code frequencies generatedand then export for future analysis. Moreover,memos can be written for adding advanced information regarding codes. It automatically generates single file out of study file. It is in- expensive and interface is easy [32].

4.3.5 Transana

It provides advanced tools for qualitative data (audio, image and video) analysis. It offers two modes of demonstrating analytic importance of client's data units, coding and categorization. In powerful means, users can use these analytic systems just as users work to generate comprehension of their data which is qualitative. Transana's text and graphical based reports are immensely customizable and flexible and allow users to discover analytic links inside data. Some features of Transana are: in single analysis, it incorporate audio, video, text and image based data; handle multifaceted media data in powerful modes; by the usage of multi-user version and cooperate with friends in real time across distance.Within Transana, users can examine data and can import text based documents. Primary aim of Transana is to simplify management and analysis of digital based data. It produces different types ofreports which is useful in delivering completevision of data and could be exported for future analysis into other tools (spreadsheet etc.) [53].

4.3.6 SPSS

Statistical package for social sciences (SPSS) software execute quantitative analysis. This software write and read data from other spreadsheets, statistical packages and databases. SPSS software only execute statistical operations. It is mostly applied in social science areas like psychology. In psychology field, techniques like t- test and cross-tabulation etc. presented within 'analyze' menu. SPSS software GUI has two sortsof views like data and variable view. Data view displays spreadsheet sight of variables and cases (columns and rows). Variable view allows user to modify it via data kinds and comprises of various headings such as type, label, name, decimals, align, measures, width and missing. SPSS limitation is that not applicable for analyzing huge set of data. In nursing and medicine fields, researchers work with huge data sets, so SAS software used in both areas rather than SPSS for analyzing clinical data. This software covers huge range of statistical techniques like data summarization (standard deviation and calculate means), determine associations between variables (regression and correlation) and graph outcomes (line graph and bar charts) [62].

4.3.7 SQL

SQL stands for 'Structured query language'. It's a programming language. Suitable for data management which resides inside relational databases. To handle data that is structured, SQL is feasible and for analysts behave like a database instrument. Applicable in data science domain. The aim behind is, mostly data kept into relational databases, when someone require to attain and need to expose its value, SQL helps out to performsuch operations. For decision making, data analysts make use of SQL to modify, access, readand analyze data which is kept inside databases. Diverse types of SQL based RDBM systems presented. Few of them are: PostgreSQL, Oracle, MySQL and MS SQL etc. There is a wide range of data visualization tools for SQL. Some include: metabase, tableau, chartio, google data studio, andlooker etc [24].

4.3.8 SAS

SAS is said to be the data analysis software. Itprovides variety of predicting techniques like: event modeling, scenario planning, what-if analysis and hierarchical

reconciliation. SAS choose variables which utilize for modeling method and creates forecasts for future. Statisticalanalysis system is a software package which designed to perform different operations such as mining of data, data modification, business intelligence and data management etc. It repossessdata from several sources and conduct statistical analysis on that data. To the feasibility of non- technical consumers, SAS provides graphic interface which is said to be point and click interface. SAS programs consist of PROC and DATA phases. PROC phase used for data analysiswhile DATA phase performs manipulation and data retrieval operations. Every phase involve chain of statements. DATA phase subdivided intotwo phases: execution and compilation. Compilation phase discover syntax errors and execute declarative statements while phase of execution emphasizes on the linear execution of executable statements. Sets of data arranged in table forms. Rows are said to be observationswhile columns are known as variables. PROCphase comprises of PROC statements. PROC statements uses named procedures. Procedures basically execute analysis task and reporting on sets of data to generate graphics, statistics and analyses. PROC based statements suitable for showing results, data sorting and for other tasks. Over three hundred named procedures exist. Every named procedure include significant form of statistical work and programming. SAS data available in different layouts like PDF, RTF, Excel and HTML etc. SAS software package include beyond two hundred modules. Few SAS modules are: quality control, data mining, statistical analysis, graphics and presentation and applications facility etc [66].

4.3.9 Python

Python is a general purpose and high levelprograming language. Python syntax is easy and simple. For data cleaning, data modeling and for creating analysis algo's, python language used in data analysis process. Main feature of thislanguage is that, it is user friendly. Moreover, it isportable. Users only write code in python language and run it into different platforms without any changeability. Programmers do not require to memorize system architecture and neither handle memory. Wide range of components, libraries and packages make python language as a useable language and many companies like Netflix, Spotify, Reddit andDropbox use python to perform operations.Furthermore, with machine learning and text mining features, python play a necessary role in advanced analysis procedures. Some popularvisualization packages are: Plotly, Seaborn and Matplotlib. Python interface also utilized for otheranalytics based systems [54].

4.3.10 R

R is said to be computing environment which emphasizes on the graphics based data representation and statistics. Moreover, it is an open source programming language interface. It comprises of more than fifteen thousand source packages, most of them used for data manipulation, data visualization, data loading and data modeling. R environment permits analysts with programming proficiency to perform and generate any kind of analysis on data. It is the high level data analysis tool, stated to as language and statisticians designed it. It can run on several platforms such as Linux, Macintosh OS, UNIX and windows etc. It is suitable to perform differenttypes of statistical base analysis like factor cluster,regression and conjoint analysis etc. R language iseasy for beginners who have no skills of high level programming language. Via usage of single

command, R execute complex mathematics problems. In statistical area, R language with widerange of graphics libraries like Plotly, shiny, Knitr, XGBoost and ggplot make R different to others. Now a days, R is widely used in companies and business industries like Facebook, google, Airbnb and twitter etc [35].

4.3.11 MATLAB

Matlab (matrix-laboratory) is a programing language as well as analytical platform. All aroundthe world, large range of scientists and engineers make use of Matlab for designing and analyzing systems. For expressing computational mathematics, matrix oriented Matlab is preferable.Integrated graphics make this language easier for data visualization. Desktop environs providediscovery, experiments and explore operations. InMatlab, analysts can perform analysis on huge setsof data. Primary feature of Matlab is that, its codecan combined with other languages that leads to employ applications and algo's in business, web and manufacturing systems. Without knowing about data layout and source, Matlab analyze massive amount of complex data in finer ways. Matlab analyze different kinds of data. Some data types are: signals, time-series, images, geospatialand dates and times [41].

4.3.12 Java

It is an object-oriented programming language.Over three billion devices execute java. It helps programmers to write once, run anywhere (WORA), it means java code is executable on variety of platforms like windows, mac and Linuxetc. Java syntax is same as C++ and C languages.It is suitable in data science domain and dataanalysis processes like import and export of data, cleaning of data, natural language processing, visualization of data and statistical analysis. For the necessity of data engineering, numerous social applications based on java which includes twitter, Facebook and LinkedIn etc. It contains various libraries and instruments accessible for machine learning and data science domains [14].

4.3.13 Datapine

It is most common business intelligencesoftware. It provides advanced analysis features for whom that require accurate and speedy onlinedata analysis solutions. By using this tool, analystscan easily combine diverse types of data sources, can run advanced analysis operations and can generate interactive dashboards. With the help of data analytics software, no one need to manually analyze huge data sets. Datapine consists of advanced level of predictive analytics features. Some features of datapine are: data alerts, drag- drop interface, self-service analytics, modern dashboards and ad-hoc queries etc.

4.3.14 Data Visualization Tools

Data visualization play a vital role in analysis process. It is the graphics based demonstration of data. Visualization tools basically provide a feasible mean to understand and to see data patterns and trends. Researchers and analysts, need to design and develop visual illustration of data. When analysts work with large volumes of data sets, procedure of making visualization makes analyst task simpler and easier. A lot of data visualization tools are available. Few of themare: charts, graphs, tables, maps and world cloud etc.

Charts

For organizing huge range of data, charts are suitable. Charts are pictures and tables. People utilize charts for interpreting present data and for making future predictions. It is said to be thegraphical data representation. In chart data represented via symbols, like lines in line chart, slices in pie chart and bars in bar chart. It can alsorepresent functions, numeric data and givesinformation. For understanding huge data volumesand associations among data parts, usually charts utilized. Charts applied in different domains and can be made via the help of computer utilizing chart base applications and by hand. Differenttypes of charts available for data representation. Few of them are: line, bar, pie, column and histogram etc.

Graphs

Graph is a kind of chart used to illustrate association among diverse data sets via devising vertical and horizontal axis. Graphs are subset of charts. Graphs utilized for raw data and show visual representation of variations and trends within data over time interval. Graphs signify mathematical relations or interconnections among different data sets. There are different types of graphs such as bar and line graphs etc. and useable for various purposes.

Word Cloud

A word cloud is a cluster or group of words portrayed in various sizes. Also said to be text or tag clouds. It is a word visualization which shows most useable words inside text from tiny to huge, in accordance with how frequently each seems. To perform word cloud, first phase is the extraction of words. It's a good concept to remove punctuation and plurals, correct spelling errors and automatically associate words which are same. Various text based analysis systems are available that require word clouds. For instance opinion mining, investigate and patent analysis [70, 82]. For the visualization of text data, word clouds are feasible and easy in utilization.

4.4 Research Techniques

Research term initiated from old French word "recerchier". It means search and search again. Incommon dialect, research denotes search for knowledge. It can be described as a systematic andscientific search for appropriate information on a particular area. Research is basically a skill of scientific exploration.

[40] state research as a structured determination to attain new knowledge.Few people deliberate research as a movement, movement from identified to unidentified.Actually itis a journey of discovery.

It is a scientific method of answering a research query, resolving an issue and creating new knowledge by methodical collection, analysis and organization of information with crucial aim of making research beneficial in decision making. Systematic research within any domain of inquiry contains three operation such as data analysis, collection and analysis of data.

In computer science research different data collection and analysis techniques are available. Few research techniques are: qualitative research, non-empirical research, qualitative research and engineering research techniques [21].

4.4.1 Quantitative Research Techniques

Quantitative research is the way of analyzing and gathering numerical based data. It could be utilized to identify averages and patterns, test causal association, make predictions and simply outcomes to broader populations. It's generally applied in social and natural sciences such aseconomics, biology, marketing, chemistry, sociology and psychology etc. Quantitative research techniques can be used for forecasting, laboratory experimentation and fieldexperimentation research.

4.4.2 Qualitative Research Techniques

Process of analyzing and gathering non- numerical based data to comprehend experiences,concepts and views. This type of research can be utilized to collect in-depth visions into a problem and produce new concepts for research. Qualitative research is generally applied in subjects of humanities and social sciences like education, history, anthropology and health sciences etc. Qualitative research techniques are applicable in action, descriptive, ethnographic andfocus group research.

4.4.3 Non-empirical Research Techniques

It's a method which contains theory development as opposed to utilizing experimentation and observation. This type of research finds solutions to problems via usage of present knowledge as its source. Although, it never means that new inventions and concepts cannot be found inside established and current knowledge pool. It plays significant part in scientific community by providing a place to startprocedure of creating scientific evidences in new research areas. Non-empirical techniques involves theorem proof, review of existing research, conceptual research and future research.

4.4.4 Engineering Research Techniques

Engineering not only concerned with knowledge of natural phenomena but concerned with how knowledge assist human requirements. Such variables like user compatibility, safety, costand adaptability to several external operating environs and situations must be taken into accountin development, designing, maintenance and operational support of services and products that engineers produce. Hence engineering contains incorporation of knowledge, methods, experiencesand techniques from various areas.

4.5 Data Analysis Techniques

In previous sections, discussed different analysis methods. This section describes differenttechniques of data analysis in terms of qualitativeand quantitative. Here we will see a briefdescription of different techniques which mainly focused on qualitative and quantitative dataanalysis.

4.5.1 Qualitative Data Analysis

Qualitative Data Analysis (QDA) is consideredto be the procedure of cataloging, depiction and linkage of phenomena along with researcher's thoughts and ideas. Phenomena beneath study must be defined accurately. QDA is said to be an iterative procedure, specify collection, analysis and processing of data operations are entwined and not succeeding procedure [61]. QA converts data into findings. It comprises of

reduction of huge amount of raw data, discover useful data patterns and develop a framework. In QDA, three steps need to be follow: description, organizing and interpretation of data [65].[80] QDA process consists of six stages: data discovery and data defining, data gathering and data storage, data sampling and reduction of data, data coding and structuring of data, testing and the last is to write up research. Researchers must be proficient in data interpretation and explanation, so, there is a need of designing and developing a conceptual framework and data need to be classified. Although for quantitative data analysis, there are some rules which need to be follow but for qualitative analysis, there are no clear rules. Some techniques for qualitative data analysis introduced, these techniques helpful for researchers from different perspectives [7].

Theoretical Prepositions

Theory is a proper, testable description of events, contains the details about how things associate with each other. Moreover, researcher describe the term preposition. "Prepositions are said to be statements deal with associations between theories". Preposition describes logical connection between thoughts via declaring a universal link among theories [86]. In scientific method, prepositions play necessary part. Preposition is same as hypothesis, but the main aim of preposition is, recommends a connection among two theories but that connection never tested with the help of experiment. So, preposition fully based on previous research. Preposition never based on testable data, so there is more difficulty to negate in scientific environment. In researcher's point of view, for example, if stated reaction is a result of an act, then this study comesunder theoretical prepositions. Researcher define,if certain preposition is wrong and if variations exists among conflicting concepts. The same way,records made for breaking, theories made for testing [86].

Triangulations

It means, researchers gain data from several sources and apply different types of techniques [7]. From doing this, researchers attain knowledge by using different techniques [30]. In triangulation, researcher utilizes one measurement method then matches findings of that method with that one, utilizing other method. If other method proves findings of first method, then triangulations reinforce finding reliability. If it never happen, then it means usage of one measurement method is not considered to be reliable [11]. According to [23], triangulation technique is expensive in termsof effort. Researcher [20] presented four kinds of triangulation. Theoretical, data, methodological and investigator triangulation. In methodological triangulation, via utilizing two methods data gathered. In investigator triangulation, data collected and analyzed via the help of two researchers. In data triangulation, through diverse type of sampling approaches, data collected. In theoretical triangulation, with the help of two theoretical based position data elucidated.

Grounded Theory

Theory which obtained from data, gatheredsystematically and examined via research procedure. Theoretical development is the primarygoal of grounded theory. In grounded theory technique, collection of data, data analysis and ultimate theory has direct association with each other. Researchers presented two main features of grounded

Table 2. Coding variations between three content analysis approaches.

Content analysistype	Study begins with	Timing of defining keywords or codes	Keywords or codes source
Summativecontent analysis	Keywords	Keywords recognized during and before analysisof data	Keywords extracted fromliterature review and researchers Interest
Conventional content analysis	Observation	Codes definedduring analysis of data	Codes extracted fromdata
Directed content analysis	Theory	Codes defined during and before analysisof data	Codes extracted from related research findings and Theory

theory. One of them is that, this technique is recursive. Recursive means, analysis and collection processes work in a circle and otherone is that it's concerned with the theory development out of data [7]. All around the world,grounded theory is considered to be the most popular research design. It begins with broad and clear research questionnaire. Then questionnaire discovers area for study. Researchers presented a list of different methods of analysis and gathering of data: data classification and initial coding, producing simultaneous data or data analysis and data gathering, intermediate coding, memos writing, theoretical integration, sampling, sensitivity and saturation, selection of primaryclass and comparative analysis through the usage of abductive and inductive logic [57]. Data isobtainable from different ways such as videos, observations, books and documents etc. to ensureanalysis validity, all type of data could be coded [16]. When in a study apply grounded theory and its approaches scholars can easily describe procedure and scheme linked with phenomenon [57]. Researchers define eleven rules which must be followed throughout analysis and data gathering for ensuring grounded theory application reliability. Meanwhile, data analysis and data collection are interconnected procedures,first of all data gathered, then it takes time in analysis due to future interviews and observationsinfluences. Concepts are important part of analysis process, because in the whole procedure of analysis, researchers require data conceptualization to work with that data. Moreover, there is a need to develop categories and must be linked with each other. Difficult for researchers to memorize all hypothesis and categories, so 'writing memos' play a vital role ingrounded theory to monitor all aspects of study. Systematic and sequential way of analysis and data gathering as biggest privilege of grounded theory [18]. Grounded theory method takes time [7].

Content Analysis

Content analysis [44] is a research based technique which is used to make proper inferencesfrom texts. It performs well when to deal with datalike open-ended surveys, interview based data anduser feedback etc. Data which need to be analyzedcould be in any form like recorded form or in written form.[71] Content analysis is a dominant technique of data depilation. The main goal of thistechnique is that, it is replicable and systematic approach which compresses text words intosmaller groups depending upon

explicit coding rules. It deals with huge data volumes [7]. For different type of unstructured data, this technique is suitable [36]. Presented three approaches ofcontent analysis. Main difference between approaches are schemes of coding and coding origins.

Framework Analysis

Framework analysis technique playing an important role in heath domain. It gives evident and systematic phases to analysis procedure, in such a way, researchers have clear perception about phases through which outcomes can be acquired from data. Framework analysis consists of five phases. These are: familiarization, charting, indexing, thematic framework identification and interpretation and mapping. In familiarization phase, partial and complete understanding of data included. Discovery of thematic framework, it is primary coding framework. Designed from emerging and theoretical issues from the phase of familiarization. In indexing phase, apply framework to data, with the help of textual and numerical codes for the discovery of particular data chunks. In the phase of charting, get headingsfrom framework for creating data charts, in such way someone can simply read complete sets of data. In the phase of interpretation and mapping, analysts find relationships, explanations, concepts and patterns of data supported with the help of plots and visual demonstrations. Analysts design plans, generate topologies, define concepts, provide clarifications and discover links inside data at this stage [68]. Researchers provide varietyof data display concepts which helpful to explore data in framework analysis terms [47,56].

Discourse Analysis

Discourse is considered to be the complete system through which individuals communicate, called 'language'. It contains both type of communication (verbal and non-verbal, written). Discourse analysis is all about examining language inside its social perspective. This technique is different from other qualitative analysis techniques that effort to comprehend meaning of societal actuality for artists in that it strives to reveal the mode inside which that realitygenerated [25]. It's an analytic technique instead of theory. It has many disciplinary roots such as anthropology, philosophy, sociology, communication studies, literature and linguistics etc [29]. Researcher identified three types of discourse analysis. Critical, formal linguistic and empirical [34]. On any data set such as (written and spoken) this analysis can be conducted. It contains focus groups, interviews, speeches and other form of texts.

Ethnographic Analysis

It contains discovery of categories associated with population, human life specially education, health related issues, family and atmosphere [55].
This analysis technique utilizes an iterative procedure [77]. In an iterative process cultural concepts rise are translated, converted and characterized in written form. Nurses applied ethnographic techniques in health domain for discovering and recording dissimilarities in how diverse type of culture and social communities accept and comprehend disease and health.

Constant Comparative Analysis

Constant comparative analysis technique takesone data piece (interview, theme and statement, etc.) and match that data piece with others that canbe same or diverse, for

Table 3. Comparison among qualitative analysis techniques.

	Ethnographic Analysis	Grounded Theory	Framework Analysis	Discourse Analysis	Content Analysis
Research question and Aim	Involvementin natural setting to attain insider capability e.g. uptake of medication, service estimations	From empirical data, generate theory e.g. stigma in mental health; views and perceptions of mental in diverse tribal groups	Mostly utilized for problem based approaches in health based services research e.g. what are the training needs for primary carestaff?	Capture variations of text or Public Discourse e.g. political theory un- derstanding, Social changeabil- Ity	Capture the meaning at Descriptive level e.g. why do carers attain facilities for Their families withanorexia?
Methods and sampling	Observational studies	Unstructured question- naire Theoretical sampling	Convenience sampling Interviews with semi- structured question- naire	Theoretical/ Purposeful sampling Speeches, newspapers, documents	Purposeful sampling Documentse.g. newspapers
Analysis	Data driven; but no fixed commitmentto developing new theory	Iterative and constant comparative approach Data driven	Deductive approach Theory driven	Complete, in- depth analysis of discourses, written text, conversa- tions, speeches	Deductive approach
Researchers position	Researchers skill and neutral position vital	Researchers position or potential bias is managed	Researcher neutral position, limited in- terpretation	Expected high level of abstraction or interpre- tation	Researcher neutral position

developing conceptualization of feasible connections among several data chunks. According to [27], this technique allocates codes which reflect methodological associations. Constantcomparative method consists of four phases: compare incidents appropriate to each group, theory delimiting, theory writing and integrating groups and properties.

Narrative Analysis

It utilizes texts like conversations, stories, photos and interviews etc. Narrative analysis behave like a device in domain of arts, education,cognitive science and sociology etc. it concerned with stories content and stories structure presentedin verbal and written form [4, 19]. Narrative analysis technique utilized for content analysiswhich collected from several means like surveys, field observations and interviews etc. Mostly stories and views shared through peopleconcentrated on the discovery of answers to research queries [77].

Sentiment Analysis

It is said to be the systematic analysis of expressions. Specially, it emphasizes on examining conducts and concepts on area of interest via usage of machine learning

approaches.In data mining domain, sentiment analysis could be explained from two ways: operational and functional. Functional way emphasizes on the practical usage of technique. For example, sentiment analysis is a strategy which classifies body of documented information for defining conducts and emotions regarding certain problem. This SA definition, describes the way of working. Operational way emphasizes on the method operations just as subarea of computational linguistics. [45] Sentiment analysis is same as opinion mining which emphasizes on categorizingand extracting texts with the help of computer programs and machine learning. In simple words, it is a data mining approach which utilizes computational linguistics, text analytics and NLP for the identification and extraction of content of heed from textual data. Sentiment analysis is the most useful and applicable qualitative technique. The goal behind it, classification and interpretation of emotions delivered inside documented data. Sentiment analysis integrates several activities to generate knowledge from textual data, in processing terms. This processing classified into five phases: collection of data, text preparation, sentiment classification and detectionand output demonstration. In data collection phase,sentiment analysis gain benefit from user created content on internet. Data can be in different forms, like emotions and behaviors etc. So, for extractingand classifying data, NLP and text analysis operations utilized. After data extraction, data willbe arranged for analysis purpose. Text preparationphases contains data cleaning which is extracted from data collection phase before performing analysis. In text preparation, elimination and identification of non-textual content performed and removed irrelevant content from textual dataset. In sentiment detection phase, with the help of computational tasks, reviews, beliefs and concepts extracted from textual dataset. Sentimentdetection performed at diverse levels such as phrases, complete sentences and documents and single term with the help of different techniques like: negation, lemmas and opinion words etc. Sentiment classification phase, classifies every particular sentence into classification sets. These sets signified on two utmost points on a scale (positive and negative, like and dislike good and bad). Classification consists of various points same to star rating utilized via retailors, hotels andrestaurants. Fifth phase show output. Several waysavailable for output presentation, but the most common of them is the usage of graphical based displays. Output presentation format based on research requirements.

4.5.2 Quantitative Data Analysis

Quantitative data analysis is measureable in number forms. Quantitative data analysis techniques concentrate on numerical, mathematical and statistical analysis of large sets of data. It contains statistical data manipulation with the help of computational algorithms and techniques. Quantitative analysis techniques mostly utilized for the explanation of certain phenomena and for making predictions. [43] Researchers provide a list of similarities among quantitative and qualitative data analysis techniques. According to researchers, both methods are same in four ways. These include: tryto escape from errors and wrong conclusions, discovery of aspects and patterns which are same or diverse, reveal study design in a useful way andmake usage of reasoning to reach on some conclusion. Different quantitative data analysis techniques are available. Some of them are as follows.

Regression Analysis

For describing biological phenomenon, regression term invented in 19th century by Francis Galton. It is quantitative data analysistechnique that utilized when researchers require modeling and analysis of variables and relationship among variables. Regression analysis basically examine association among dependent (known as outcome variable) and independent variables (said to be features, predictors and covariates). The goal behind is to evaluate how one or more variables effect on dependent variablefor the identification of patterns and trends. It is feasible to make predictions regarding future trends. This statistical technique also utilized to comprehend which independent variables are connected to dependent variables and help out to discover relationships. Regression analysis could be utilize to understand casual relations among variables but this may actuate to wrong associations. Several kinds of regression models are available but selection of right model based upon data type. Before performing analysis, every regression technique consists of few assumptions which must be fulfil. These techniques vary regarding the kind of independent and dependent variables.

Regression Analysis Types: In medical domain, usually three kinds of regression analysisused. These are linear, cox and logistic regression.Linear regression is the easiest model to use because its demonstration is easy. A technique in which dependent variable is said to be continuous. Connection among dependent and independent variables supposed to be linear. For predictiveanalysis, linear regression is the most useable technique. Linear regression is sub-divided into two types: simple and multiple. When workedwith the help of one dependent and one independent variable then it is said to be simple linear regression. On the other side, when used one dependent and more than one independent variable known as multiple linear regression. It isfeasible when dependent variable is said to be binary in nature but if it is non-binary then multinomial or polynomial logistic regression canbe used. Logistic regression not require linear connection among variables. Special type of regression and feasible for 'time-to-event' data. For instance, time from first heart attack to second[28].

Monte Carlo Simulation

It is also said to be Monte Carlo method. It is acomputerized method suitable for producing models of viable outputs and their probability dispersions. Basically it considers variety of viable outputs and evaluates how specific output would be comprehended. This method is feasible for data analysts for conducting advanced level of risk analysis which allow data analyst to forecast about what can happen in future and take decisions.It is the commonly used method for estimating influence of random variables on a particular outcome variable which makes this method suitable for risk based analysis [67].

Factor Analysis

A technique which reduces huge range of variables to fewer range of aspects. Its working depends on that several distinct, evident variablesrelate with one another because all connected withfundamental set up. It is feasible for two reasons. One is it reduces huge sets of data into fewer one and second is it expose concealed patterns. [15] FA utilizes mathematical processes for identifyingpatterns from variables set. [84] Primary aim of factor analysis is the data summarization in such away that patterns and

associations could easily beunderstand and interpreted. It is applicable in numerous domains like medicine, geography, social and behavioral sciences and economics. It permits researchers to explore thoughts which could not be perceive and quantify easily like joy,wealth and satisfaction etc.

Types of Factor Analysis It can be divided into two types: exploratory and confirmatory.

Exploratory Factor Analysis Main goals behind exploratory factor analysis are: determine the quantity of common aspects effecting variables set and to determine the power of association among every aspect and every perceived measure. It is suitable to expose the underlying arrangement of relatively hugevariables set. It attempts to reveal complex patterns via discovering sets of data and predictions testing [15]. Some computer programs like SPSS 'statistical package for social sciences' and SAS 'statistical analysis system' permit for the usage of exploratory factor analysis. Althoughevery program offers diverse type of techniques and options but some phases are same like rotation,extraction, interpretation and selection [73]. Via usage of two methods (Q and R type factor analysis) exploratory factor analysis can be done. In Q type analysis, factors evaluated from single respondent while in R type analysis, factors evaluated with the help of correlation matrix.

Confirmatory Factor Analysis It tries to approve hypothesis and for the representation of factors and variables utilizes path analysis diagrams. It works just as a tool which is applied for the approval and rejection of theory of measurement. Two sorts of model need to be explored in confirmatory factor analysis: structural equation and measurement model. For several areas of behavioral and social sciences, CFA considered compulsory tool of analysis. CFA comes under the family of structural modeling approaches which permit for examining casual associations between observed and hidden variables. Primary objective of CFA is that it helps researchers to observe slot among observation and theory. SEM behave like an umbrella and it comprises of several techniques for data analysis. Some statistical analysis examples found comes under the SEM umbrella like multi-group confirmatory factor analysis, confirmatory factor analysis, for longitudinal analysis-latent models and regression with the help of latent variable outputs. Confirmatory factor analysis is an important phase to perform SEM model types [10]

Cohort Analysis

Term cohort defines cluster of people. Cohort study is referred to as research paradigm focusingon a specific cluster with same characteristics andperceives over time interval. Cohort study can be classified into two types: prospective and retrospective cohort study. Prospective cohort studies are said to be scheduled earlier and accomplish over future division of time.Retrospective cohort studies examine data thatavailable and attempt to discover risk factors for specific situations. Elucidations are finite because scholars cannot collect lost data and never go back[52]. It is a kind of behavioral analytics which accesses data from dataset and instead of observing all users just as single component, for analysis it splits them into associated sets. These cohorts or associated sets basically share mutual experiences and attributes under a fix time period. Cohort is basically a set of individuals who share mutual features. For instance, students who joineduniversity in 2020, denoted to as cohort 2020. For analyzing client conduct, corporations apply cohort analysis. Without this, companies face problems to

understand life process that every client experience over a certain time interval. Companies perform cohort analysis for understanding patterns and trends of clients over time. There are some phases which helps to execute cohort analysis properly. These phases are: define what query you need to be respond, determine metrics which will be helpful to response against query, determine particular cohorts which relate with each other, perform cohort analysis and the last phase is results testing[52]. There are three kind of cohorts: time, size and segment oriented cohort. During a given time interval, time oriented cohorts are those whenclient registered for a service. Cohort analysisdepicts client conduct based on the time when clients begun to use company service and product.Segment oriented cohorts are said to be thoseclients who bought and paid for a particular product and service in past. It clusters clients by product kind and service level they registered for.Client needs might be different according to sign up for basic and advanced level services. Size oriented cohort denote numerous sizes of clients who buy company service and product. Clients might be enterprise level, small and middle sized businesses.

Time Series Analysis

It is a statistical based technique utilized for the identification of cycles and trends over time. Timeseries data is an arrangement of data points that measure identical variable at diverse time intervals.Time series is said to be the gathering of observations made consecutively via time. Time series contains single variable records is said to be univariate while time series holds more than one variable records known as multivariate. Time series could be in two forms: continuous and discrete.Discrete series involves observations which measured at separate points of time whereas continuous series contains observations which measured at every instance of time [1]. Several forces and causes which influence observation values within time series are modules or components of time series. Four components are: seasonal, random, trend and cyclic modules. Trend illustrates the common tendency of data to reduce or enhance throughout extensive time period. It is noticeable that propensities may enhance, reduce or stable in diverse time units. Cyclic changes defines medium term variations intime series, affected via situations which replicate within cycles. Cycle duration extends over extensive time interval, commonly more than one year. For instance, business cycle involve four steps: depression, decline, recovery and prosperity.Seasonal changes are variations in time series happened within a year throughout season. Primary aspects causing seasonal changes are: weather conditions, customs and climate etc. For producers, and businessmen, seasonal variation play a vital role for developing future plans properly. Irregular variations initiated via uncertain effects, which never reprise in a certain pattern and also not regular. So random variations happened due to flood, war, earthquake and strike etc. No statistical technique available to measure random variations [1].

Cluster Analysis

Several models of data mining based on some notion of resemblance among information pieces encrypted in concerned data. Numerous names suggested for methods of clustering, relying on thearea of application in data science. For instance, ‘numerical taxonomy’ term applied in biology[76], ‘Q analysis’ term often applied in psychology, ‘unsupervised pattern recognition’ term is applicable in artificial

Table 4. Comparison among qualitative and quantitative data collection methods.

Qualitative Data Collection Methods	Quantitative Data Collection Methods
Unstructured or semi-structured Techniques	Fixed and more structured techniques
Not test oriented or instrument	Test oriented or instrument
Not measureable	Measurable
Not utilized for statistical based Tests	Utilized for statistical based tests
Small sample size	Large sample size
Data generally ordinal or Nominal	Data generally interval or rational
Text based to gather Information	Number based to gather information
Interview forms contain open-ended questionnaires	Interview forms contain close-ended questionnaires
Lack of strong scientific control	Applies strong scientific control

intelligence research and 'segmentation' term used market researchers. 'Cluster analysis' term applied as a step in the process of knowledge discovery [78]. Cluster analysis discover structures inside set of data. Theobjective behind is the sorting of diverse datapoints into clusters internally similar. In simple words, data points inside a group are same with one another and different to data points inside another group. Clustering is useful to understand how data dispersed in a certain set of data or for other algorithms just as preprocessing phase. For statistical analysis of data, cluster analysis is the most popular technique and highly useable in various domains like information retrieval, bioinformatics, pattern recognition, computer graphics, machine learning, data compression anddata mining. Clustering methods is sub-divided into two sets: partitional and hierarchal clusteringmethods [49]. Partitional methods concurrentlysplit group of data points into sub-groups while hierarchal cluster methods find consecutive clusters by utilizing previously conventionalclusters. Different types of clusters are available. These are: centroid, connectivity, distribution and density clustering. Clustering has huge quantity ofapplications spread over several fields. Some popular clustering applications are: image and market segmentation, anomaly detection, social network analysis, medical imaging and search result grouping. Clustering is said to be an unsupervised machine learning technique but it can be utilized for improving accuracy ofsupervised machine learning algorithms.

5 Conclusion

Evaluation in every research work focuses on information collection and then analyzing different scenarios. Success and quality of research work is based on the knowledge of data collection application and methods. Data type is significant in statistical analysis and appropriate knowledge of is compulsory for analyzing sets ofdata with suitable selection of methods. Qualitative and quantitative information falls in acontinuum and differs according to the kind of data, collection method, tools and approaches of data analysis. This paper presented diverse kinds of data collection methods which can be applied inany evaluation along with complete depiction of numerous techniques and

tools employed by researchers for data collection. Starting with the explanation of secondary and primary data,quantitative and qualitative data and every data collection method is defined in detail. Various methods which have been described with focus on groups, observation, case study, experiment and interview etc. Moreover we showed holistic view of different types of qualitative and quantitative tools which can be used for analyzing data. Numerous statistical analysis software packages are available like SPSS and SAS etc. With the helpof these packages, it becomes simple and easy to obtain output and interpret results. Researchers must interpret diagrams, tables and statistics accurately. The main purpose of this paper is to differentiate and elaborate the different quantitative and qualitative techniques in computing discipline.

References

1. Adhikari, R., Agrawal, R.K.: An Introductory Study on Time Series Modeling and Forecasting. arXiv preprint (2013). doi: 10.48550/ar.
2. Oluwatosin Ajayi, V.: Primary Sources of Data and Secondary Sources of Data. Benue State University (2017)
3. Antonius, R.: Interpreting Quantitativedata with SPSS. Sage (2003)
4. Bamberg, M.: Narrative Analysis. APA Handbook of Research Methods in Psychology, 3 (2010). doi: 10.1037/13620-006.
5. Baskarada, S.: Qualitative Case Study Guidelines. The Qualitative Report, 19(40), pp. 1–25 (2014). doi: 10.46743/2160-3715/2014.1008.
6. Bazeley, B.: Qualitative Data Analysis with Nvivo, pp 6–15 (2007)
7. Bell, E., Bryman, A., Harley, B.: Business Research Methods. Oxford University Press (2018)
8. Brewer, J.: Ethnography. McGraw-HillEducation (2000)
9. An Introductory Brief.: Qualitative and Quantitative Research Techniques for Humanitarian Needs Assessment. Physical Review, 47, pp. 777–780 (2012)
10. Brown, T.A., Moore, M.T.: Confirmatory Factor Analysis. Handbook of Structural Equation Modeling, pp. 361–379 (2012)
11. Bryman, A.: Integrating Quantitativeand Qualitative Research: How is it Done?. Qualitative Research, 6(1), pp. 97–113 (2006). doi: 10.1177/1468794106058877.
12. Burchfield, R.W.: The NewFowler's Modern English Usage. Clarendon Press Oxford (1996)
13. Checkland, P., Holwell, S.: Information, Systems, and Information Systems. John Wiley & Sons Chichester (1998)
14. Chekanov, S.V.: Scientific Data Analysis Using Jython Scripting and Java. Springer Science & Business Media (2010). doi: 10.1007/978-1-84996-287-2_1.
15. Child, D.: The Essentials of Factor Analysis. A&C Black (2006)
16. Corbin, J.M., Strauss, A.: Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. Qualitativesociology, 13(1), pp. 3–21 (1990). doi: 10.1007/BF00988593.
17. Costa, A.P., de Souza, F., Moreira, A., Neri de Souza, D.: Webqda–Qualitative Data Analysis Software: Usability Assessment. In: 2016 11th Iberian Conference on Information Systems and Technologies, pp. 1–6 (2016). doi: 10.1109/CISTI.2016.7521477.
18. Dalati, S.: Measurement and Measurement Scales. In: Modernizing the Academic Teaching and Research Environment, pp 79–96 (2018)
19. Demuth, C., Mey, G.: Qualitative Methodology in Developmental Psychology. International Encyclopedia of Social and Behavioral Sciences (2015). doi: 10.1016/B978-0-08-097086-8.23156-5.
20. Denzin, N.K.: Triangulation 2.0. Journal of Mixed Methods Research, 6(2), pp. 80–88 (2012). doi: 10.1177/1558689812437186.

21. Dignos, A.: Research Methods (2019)
22. Field, A.: Discopering Statistics Using SPSS (2009)
23. Flick, U.: Mixing Methods, Triangulation, and Integrated Research.Qualitative Inquiry and Global Crises, 132(1), pp. 1–79 (2011). doi: 10.4324/9781315421612-7.
24. Fotache, M., Strimbei, C.: SQL and Data Analysis. Some Implications for Data Analysits and Higher Education. Procedia Economics and Finance, 20, pp. 243–251 (2015). doi: 10.1016/S2212-5671(15)00071-4.
25. Geertz, C.: The Interpretation of Cultures. Basic Books (1973)
26. Gill, P., Stewart, K., Treasure, E., Chadwick, B.: Methods of Data Collection in Qualitative Research: Interviews and Focus Groups. British Dental Journal, 204(6), pp. 291–295 (2008). doi: 10.1038/bdj.2008.192.
27. Glaser, B.G., Strauss, A.L.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Mill Valley, CA: Sociology Press (1967)
28. Gogtay, N.J., Deshpande, S.P., Thatte, U.M.: Principles of Regression Analysis. Journal of the Association of Physicians of India, 65, pp. 48–52 (2017)
29. Grant, D., Michelson, G., Oswick, C., Wailes, N.: Guest Editorial: Discourse and Organizational Change. Journal of Organizational Change Management (2005). doi: 10.1108/09534810510579814.
30. Graue, C.: Qualitative Data Analysis. International Journal of Sales, Retailing & Marketing, 4(9), pp. 5–14 (2015)
31. Greenbaum, T.L.: The Handbook for Focus Group Research. Sage Publications (1998), doi: 10.4135/9781412986151.
32. Hesse-Biber, S., Dupuis, P., Kinder, T.S.: Hyperresearch: A Computer Program for the Analysis of Qualitative Data with an Emphasis on Hypothesis Testing and Multimedia Analysis. Qualitative Sociology, 14(4), pp. 289–306 (1991)
33. Yahmady, A., Hilal, H., Al Abri, S.S.: Using Nvivo for Data Analysis in Qualitative Research. International Interdisciplinary Journal of Education, 2(2), pp. 181–186 (2013)
34. Hodges, B.D., Kuper, A., Reeves, S.: Discourse Analysis. Bmj, 337, pp. 879 (2008). doi: 10.1136/bmj.a879.
35. Hothorn, T., Everitt, B.S.: A Handbook of Statistical Analyses Using R. CRC press (2014). doi: 10.1201/b17081.
36. Hsieh, H.F., Shannon, S.E.: Three Approaches to Qualitative Content Analysis. Qualitative Health Research, 15(9), pp. 1277–1288 (2005). doi: 10.1177/1049732305276687.
37. Hurmerinta-Peltomäki, L., Nummela, N.: Mixed Methods in International Business Research: A Value-Added Perspective. Management International Review, 46(4), pp. 439–459, (2006). doi: 10.1007/s11575-006-0100-z.
38. Husamaldin, L., Saeed, N.: Big Data Analytics Correlation Taxonomy. Information, 11(1), pp. 17 (2020). doi: 10.3390/info11010017.
39. Ibrahim, M.: The Art of Data Analysis. Journal of Allied Health Sciences Pakistan, 1(1), pp. 98–104 (2015)
40. Kabir, S.M.S.: Basic Guidelines for Research. Chittagong: Book Zone Publication (2016)
41. Knight, A.: Basics of Matlab and Beyond. Crc Press (2019)
42. Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y.: Exploratory Data Analysis. In Secondary Analysis of Electronic Health Records. Springer, pp. 185–203 (2016)
43. Kreuger, L., Neuman, W.L.: Social Work Research Methods: Qualitative and Quantitative Approaches: With Research Navigator. Pearson/allyn and Bacon (2006)
44. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage Publications (2018)
45. Kumar, A., Teeja, M.S.: Sentiment Analysis: A Perspective on its Past, Present and Future. International Journal of Intelligent Systems and Alications, 4(10), pp. 1 (2012)
46. Kumar, V., Garg, M.: Predictive Analytics: A Review of Trends and Techniques. Int. J. Comput, 182, pp. 31–37 (2018)

47. Lacey, A., Luff, D.: Qualitative Data Analysis. Trent Focus Sheffield (2001)
48. Madden, A.D.: A Definition of Information. Aslib Proceedings: New Information Perspectives. Emerald Group Publishing Limited, pp. 343–349 (2000)
49. Madhulatha, T.S.: An Overview on Clustering Methods. Arxiv Preprint Arxiv (2012). doi: 10.48550/arXiv.1205.1117
50. Marshall, C.: Roschman, g. Designing Qualitative Research. Sage Publications California (1989)
51. Marshall, C., Rossman, G.B.: Designing Qualitative Research. Sage Publications (2014)
52. Mason, W.M., Wolfinger, N.H.: Cohort Analysis. International Encyclopedia of the Social & Behavioral Sciences (2001)
53. Mavrikis, M., Geraniou, E.: Using Qualitative Data Analysis Software to Analyse Students' Computer-mediated Interactions: The Case of Migen and Transana. International Journal of Social Research Methodology, 14(3), pp. 245–252 (2011)
54. McKinney, W.: Python for Data Analysis: Data Wrangling with Pandas, Numpy, and Ipython. O'reilly Media (2012)
55. Merriam, S.B.: Qualitative Research and Case Study Applications in Education. Revised and Expanded from Case Study Research in Education. Eric (1998)
56. Miles, M.B., Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. Sage (1994)
57. Mills, J., Birks, M.: Qualitative Methodology: A Practical Guide. Sage (2014)
58. Mishra, P., Pandey, C., Singh, U., Gupta, A.: Scales of Measurement and Presentation of Statistical Data. Annals of Cardiac Anaesthesia, 21(4), pp. 419 (2018)
59. Morgan, D.L., Krueger, R.A.: The Focus Group Guidebook. Sage (1998)
60. Musante, K., DeWalt, B.R.: Participant Observation: A Guide for Fieldworkers. Rowman Altamira (2010)
61. Nieuwenhuis, J.: Analysing Qualitative Data. First Steps in Research, 2, pp. 104–130 (2007)
62. Ong, M.H.A., Puteh, F.: Quantitative Data Analysis: Choosing Between Spss, Pls, and Amos in Social Science Research. International Interdisciplinary Journal of Scientific Research, 3(1), pp. 14–25 (2017)
63. Osang, J., Udoimuk, A., Etta, E., Ushie, F., Offiong, N.: Methods of Gathering Data for Research Purpose and Applications Using Ijser Acceptance Rate of Monthly Paper Publication. Iosr Journal of Computerengineering Iosrjce, 15(2), pp. 59–65 (2013)
64. Ott, R.L., Longnecker, M.T.: An Introduction to Statistical Methods and Data Analysis. Nelson Education (2015)
65. Patton, M.Q.: Two Decades of Developments in Qualitative Inquiry: A Personal, Experiential Perspective. Qualitative Social Work, 1(3), pp. 261–283 (2002)
66. Peng, C.J.: Data Analysis Using SAS. Sage Publications (2008)
67. Platon, V., Constantinescu, A.: Monte Carlo Method in Risk Analysis for Investment Projects. Procedia Economics and Finance, 15(14), pp. 393–400 (2014)
68. Ritchie, J., Spencer, L.: Qualitative Data Analysis for Applied Policy Research. The Qualitative Researcher's Companion, 573, pp. 305–29 (2002)
69. Sapsford, R., Abbott, P.: Ethics, Politics and Research. R. Sapsford &v. Ju Eds. Data Collection and Data Analysis, pp. 317–340 (1996)
70. Stasko, J., Goörg, C., Liu, Z.: Jigsaw: Supporting Investigative Analysis Through Interactive Visualization. Information Visualization, 7(2), pp. 118–132 (2008)
71. Stemler, S.: Practical Assessment, Research and Evaluation (2002)
72. Sundaram, K.R., Dwivedi, S.N., Sreenivas, V.: Medical Statistics: Principles & Methods. Anshan (2010)
73. Tabachnick, B.G., Fidell, L.S.: Using multivariate statistics. Allyn and Bacon. Needham Heights Ma (2001)
74. al, E.R.T.E.: Data Collection Methods for Evaluation: Document Review (2018)
75. Thompson, C.B.: Descriptive Data Analysis. Air Medical Journal, 28(2), pp. 56–59 (2009)

76. Thorel, M., Krichevsky, M., Lévy-Frébault, V. V.: Numerical taxonomy of mycobactin-dependent mycobacteria, emended description of Mycobacterium avium, and description of Mycobacterium avium subsp. avium subsp. nov., Mycobacterium avium subsp. paratuberculosis subsp. nov., and Mycobacterium avium subsp. silvaticum subsp. nov. International Journal of Systematic and Evolutionary Microbiology, 40(3), pp. 254–260, (1990).
77. Thorne, S.: Data analysis in qualitative research. Evidence-Based Nursing, 3(3), pp. 68–70, (2000).
78. Thrun, M. C.: Approaches to cluster analysis. In: Projection-Based Clustering Through Self-Organization and Swarm Intelligence. Springer, pp. 21–31, (2018).
79. Venkatesh, V., Brown, S. A., Bala, H.: Bridging the qualitative–quantitative divide: guidelines for conducting mixed methods research in information systems. MIS Quarterly, pp. 21–54, (2013).
80. Watling, R., James, V., Briggs, A.: Qualitative data analysis: using NVivo. In: Research Methods in Educational Leadership & Management, pp. 381–396, (2012).
81. Whitehead, D., Annells, M.: Sampling data and data collection in qualitative research. In: Nursing and Midwifery Research: Methods and Appraisal for Evidence-Based Practice, pp. 105–121, (2007).
82. Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., Qu, H.: Opinionseer: interactive visualization of hotel customer feedback. IEEE Transactions on Visualization and Computer Graphics, 16(6), pp. 1109–1118, (2010).
83. Yilmaz, K.: Comparison of quantitative and qualitative research traditions: epistemological, theoretical, and methodological differences. European Journal of Education, 48(2), pp. 311–325, (2013).
84. Yong, A. G., Pearce, S.: A beginner's guide to factor analysis: focusing on exploratory factor analysis. Tutorials in Quantitative Methods for Psychology, 9(2), pp. 79–94, (2013).
85. Zainal, Z.: Case study as a research method. Jurnal Kemanusiaan, 5(1), (2007).
86. Zikmund, W. G., Babin, B., Carr, J., Griffin, M.: Business Research Methods, 7th ed. Thompson Learning, CA, (2003).
87. Zikmund, W. G., Carr, J. C., Griffin, M.: Business Research Methods. Cengage Learning, (2013).

ISSN 1870-4069

Audio Signal Analysis for Indexing and Rating Movies

Abdullah[1,2], Nida Hafeez[1,2], Muhammad Ateeb Ather[2], José Luis Oropeza-Rodríguez[1], Alexander Gelbukh[1]

[1] Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico

[2] Department of Computer Sciences, Bahria University, Lahore, Pakistan

abdullah2025@cic.ipn.mx, nhafeez2024@cic.ipn.mx, joropeza@cic.ipn.mx, gelbukh@cic.ipn.mx, 03-134211-022@student.bahria.edu.pk

Abstract. Identification of the contents of a video sequence is a major factor in determining the context of the video. Video classification has been a popular research area in the fields of Computer Vision, Artificial Intelligence and Natural Language Processing for long now. Classifying videos on the basis of factors like video frames, audio signals, static images and similar key factors have been worked upon. So far, techniques like Hidden Markov Models, Dynamic Time Wrapping and Neural Networks have been incorporated for content-context based classifications of videos. This paper discusses social and technical effects of a specific languages being spoken in a video and focuses on determining the suitable audience for that video sequence i.e. a movie. We have proposed a system that helps classifying the context on the basis of audio track of the video. Natural language processing plays the primary role in the proposed system. Other techniques have also been used according to the different requirements of the system.

Keywords: Natural language, speech recognition, speech synthesis context based classification, natural language processing

1 Introduction

A video sequence is a rich source of information that consists of objects, motion (movement), speech (dialogues), and text (captions), colors (B&W, Greyscale, and Colored) and images. Humans can easily and quickly interpret both the semantic and syntactic context of a video sequence from the information being provided by the video itself. With the advancements in technology, internet is available to anyone, everyone, anywhere and everywhere. With this comes a pressing need for context based

classification since the internet is accessible by children too. There is a need for more efficient and effective tools and methods for dissemination of visual and audible content available on the internet. This simply states that multimedia content needs to be indexed, classified and stored according to their context and be accessible for suitable audience only. This first and foremost requirement for context-based indexing is understanding and interpreting the context before further processing. Other processes include sorting, calculating, storing etc. the subject according to the context.

The key factor in understanding the content and context of a video scene segmentation and audio track of the given sequence. Research has focused on use of information provided by the speech and images. The most common techniques we come across include video segmentation, video frames, static images, identifying objects have been used to interpret the context of a video sequence in terms of visual features. In addition, several researchers have worked on analyzing audio signal for context-based classification. This theory is feasible because audio tracks in different environments are fairly differentiable. For example, the audio of a film is different from that of a documentary.

Although this thought may pop in mind that audio signal alone may not be enough for indexing and categorization of a video and that video analysis would be necessary. However, in this paper, we will be discussing audio-based analysis. Because it's significantly less complex and uses simpler computations. We also propose a technique that will perform the audio analysis of movies and then categorize and rate them according to appropriate audience.

The major highlight of our technique is that it uses Natural Language Processing and the dataset is updateable. Our technique comprises of three steps; first we extract the audio track of any video movie; the system then analyzes and compares the audio, converts it into text using speech recognition and is fed to the system; the system compares the text with the predefined data set; on the basis of the results obtained, simple mathematical calculations are performed and the movie can be rated for different audiences.

2 Motivation

Since the last few decades, there has been a lot of advancement and development in the technological sector. In the late 1990s, internet was released commercially and made available to everyone. Businessmen, Educators, Students all have now easy access to the internet via WIFIs, Edge and now recently released 3G in their smartphones, tablets, laptops and PDAs. Internet plays a major role in everyone's daily life. Whether its social media, news, watching programs online, viewing top charts or looking for a recipe, internet has makes everything you ask for available in the matter of a second. Young or old, everyone needs internet nowadays. But with pros comes cons. Not everything is suitable for everyone. The context of content can vary from one age group to another. What might be appropriate for one age group may not be suitable for the other. In this paper, focus has been laid on classifying movies according to different

age groups. These age groups are classified as PG-13, PG-18, and Rated-R in terms of classification with respect to the context. Following this, audience can review the rating and therefore interpret whether the movie is suitable or them or not.

3 Related Work

In the recent past, researchers have focused towards investigating the potential of analyzing an audio signal for video classification [5-6]. Saunders proposed a method for separating speech from music [9]. Nam and Tewfik presented to detect sharp temporal variations [8]. Lie et al. proposed a method to detect change in the feature vectors [6].

Saraceno and Leonardi presented a technique for differentiating silence, speech music and voice clips from an audio sequence [5] and so did Pfeiffer [7]. Lie Lu presented a theory for audio content analysis on the basis of nature of the audio using an algorithm based on the KNN method [2].

Wold, Blum and Wheaton proposed a method to classify audio content on the basis of its characteristics such as pitch, sharpness, beat and rhythm [3].

Mahedero, Martinez and Cano have proposed a theory on the natural language processing of music lyrics using language identification, structure extraction and thematic analysis [16].

Gunsel and Tekalp have discussed automatic scene change detection and key-frame selection for content-based video abstraction using automatic threshold selection techniques [17].

Ferman, Tekalp, Mehrotra have proposed a methodology for effective video content representation using temporal segmentation and domain-specific implementation such a sports, dramas/movie and news etc. [18].

Tsekeridou and Pitas have done a research on audio-visual analysis that analyzes audio and video information and interprets their relationship [19].

Liu et al. used automatic audio classification and speaker identification techniques for content-based video analysis and classification [20].

Chunneng Huang et al. proposed a framework that uses text-based content classification of videos on online video sharing sites using user-generated data. Huang et al. discussed about recent changes in performance of speech recognition systems for larger data sets. Reduction in word error rate and improvements were observed for a large corpus [22].

Wei Jiang et al. investigated the incorporation of Short-term Audio Visual Atoms for video concept classification [23]. Yuichi Nakamura and Takeo Kanade highlighted the association between the images clues and language clues to conclude corresponding video segments using natural language processing and scene segmentation [24].

Finally David J. White et al. studied the role video recognition in behavioral research and developed a system that uses voice recognition to collect and accumulate data [25].

4 Types of Speech Recognition Systems

Several speech recognition systems available on the market. A strong speech recognition system can trace over thousand words. These systems usually require training before they can be put to practical and professional use. Normally, we come across two types of speech recognition systems. Speaker-Dependent and Speaker-Independent. The third type, a more recent and therefore a less common type, known as Speaker-Adaptive.

4.1 Speaker Dependent

This system works by learning characteristics, unique attributes of the user like his voice, his accent, etc. so that the system becomes accustomed to a specific person. Users train the system by speaking to the system so that it can analyze how the user talks. Usually the user has to read a few texts pages to the system before they can use it. Speaker-dependent systems are usually easy to develop, cheap and more efficient to use, but not as dynamic as speaker independent or speaker adaptive systems.

4.2 Speaker-Independent

A speaker-independent system is developed to recognize anyone's voice and is operable for any number of speakers (of a particular dialect e.g. American English). Since it has been designed to interpret anyone's voice, it doesn't require any prior training like the speaker-dependent system. Such systems are more feasible to use in interactive environments where users can just use the system without having to read any text to the system to train it.

Having said that, speaker-independent systems in interactive environments where users can just use the system without having to read any text to the system to train it. Having said that, speaker-independent systems in contrast to speaker dependent systems are less accurate and not as efficient.

4.3 Speaker-Adaptive

Another trend in speech recognition systems is emerging nowadays, known by the name of speaker-adaptive systems. The unique feature these systems offer is that these system start off as speaker-independent systems. And later on adjust themselves according to individuals in a brief training period.

5 Speech Recognition Input Types

Different speech recognition systems have different input mechanisms. On the basis of the input types, speech recognition systems are distinguished as follow.

5.1 Isolated Word Recognition

As the name suggests, this input type system requires the user to input one utterance at a single time. This means that there needs to have quite at the beginning and after termination. It may seem that the system accepts only one word inputs but this is not the case. Rather the system requires pauses in between so that it can process the input during these pauses. Such systems have "Listened/Not Listened" states that define if the system heard the input word clearly or not.

5.2 Connected Word Recognition

Similar to Isolated Word Recognition systems, Connected Word Recognition systems require accept a combination of words to be input once at a time with a short pause in between inputting each chunk.

5.3 Continuous Speech Recognition

Continuous speech recognizers are more advanced form of speech recognition systems. That is why they are one of the more difficult to develop because they require special methods to interpret the input signal. These systems allow the user to speak naturally while the system process the content.

5.4 Spontaneous Speech Recognition

There is a number of explanations as to what is spontaneous speech. Basically, spontaneity of a speech is determined by its naturalness. A speech recognition system with spontaneous speech recognition has the ability to interpret and process natural speech expressions such as "ums" and "ahs" or other similar expressions.

6 Mechanism

The microphone converts the digital signal into an analog on. This is done by the sound card present in the computer. The user inputs the signal, also known as utterance which is a binary sequence consisting of 1's and 0's which also comprise programming languages. Sound-recognition systems comprises of two models, acoustic model and language model.

In the first step, the acoustic model (An acoustic model takes audio recordings and text descriptions to form statistical representation of the sound of each word.) which breaks down the voice signal into speech elements known as phonemes. More recent versions of speech recognition systems have matured so that now they can eliminate noise and other distracting factors in speech recognition process.

After this, in the second step (language model), the input signal is compared to the data set, the "digital dictionary", fed to the computer. This dictionary contains over

100,000 words, a fairly large collection. If the system finds a match from the digital dictionary, it is displayed on the screen.

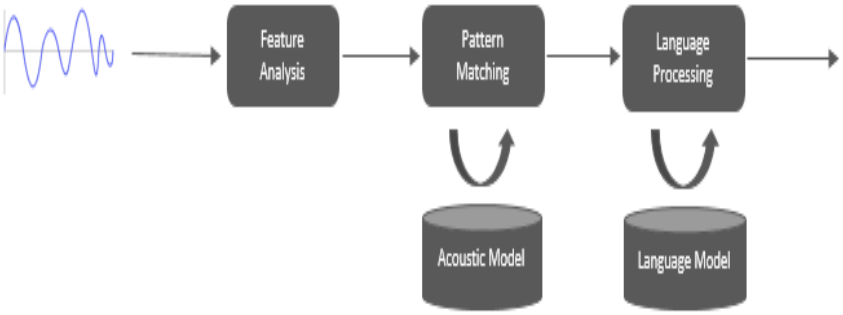

Fig. 1. General processing pipeline.

7 Algorithms used for Speech Recognition

Language modelling is an important part of modern speech recognition algorithms.

7.1 Hidden Markov Models (HMMs)

In HMMs, the state of the system is not visible but the output is visible unlike simple Markov Models in which the states are visible. Using HMMs a speech signal can be viewed as a stationary signal in piecewise. On a short time-scale, speech can be regarded as a stationary process and therefore can be regarded as a Markov Model.

Nowadays general purpose speech recognition systems are based on HMMs. Hidden Markov Models are statistical models that output the given input string in a sequence.

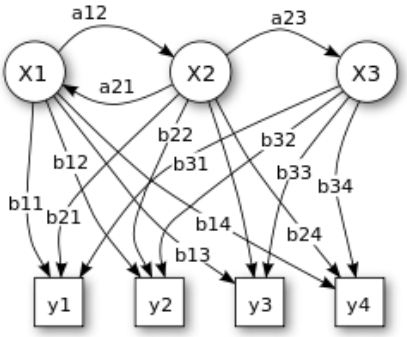

Fig. 2. HMM example.

The system being used is considered to be a Markov method with hidden states. Another reason why HMMs are popular is that they simple, computationally feasible

to incorporate. HMMs are especially recognized for their use in speech, gesture sensors, handwriting recognition, P-O-S tagging and bio-informatics.

In the recent past, HMMs have been generalized to duplets and triplets models to allow more complex data structures to be considered.

7.2 Dynamic Time Wrapping

Dynamic Time Wrapping is used for measuring similarity between two varying temporal sequences. They may vary in time or speed. Dynamic Time wrapping can be applied on any data that can be converted into a linear sequence. A popular application of DWT is speech recognition to deal with different speaking speeds.

Dynamic Time Wrapping based speech recognition systems were traditionally used but have now largely been replaced by Hidden Markov Models. Dynamic Time Wrapping is used for measuring similarity between two input sequences that may vary in time or speed.

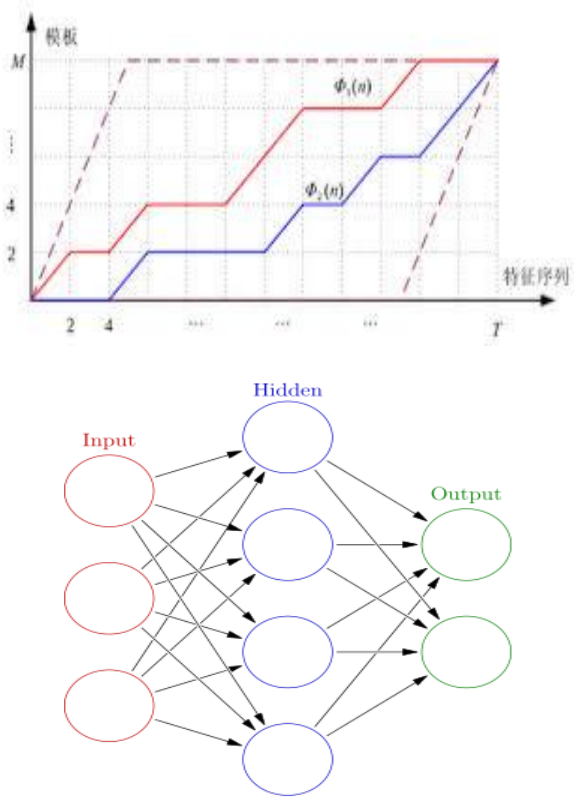

Fig. 3. Dynamic type wrappig algorithm.

One well known application of Dynamic Time Wrapping algorithm is automatic speech recognition which allows a computer to find an optimal match between two given sequences within certain bounds.

7.3 Neural Networks

Inspired from the biological neural networks, Neural Networks are computational algorithms used approximate or estimate functions that depend on a large number of unknown inputs. Neural networks emerged in the late 1980. Since then, they have been used in many different aspects of speech recognition including phoneme classification, isolated word recognition and speaker adaptation.

Unlike HMMs, these do not assume about statistical properties of the given input and possess several other qualities which makes them a better choice for speech recognition. But despite their effectiveness in classifying short-time units, they seldom succeed in continuous recognition tasks.

Like other systems that learn from the input data, Neural Networks have been applied in a number of hard tasks including speech recognition and computer vision using simple rule-based programming.

8 Audio Basic Properties

Audio just like any other entity possess properties. We will focus on two types of properties. Physical; and cognitive. Physical properties include measureable properties such as amplitude and phase. Whereas cognitive properties include properties associated with cognitive senses such as loudness and pitch.

8.1 Physical Properties

Sound is produced when the air pressure changes which is represented in the form of a wave which in turn is made up of sine waves having different frequencies, amplitude and phase. From experiments, it has been concluded that human ear does not sense change in the phase but does recognize change in the amplitude such as loudness and also any changes in the frequency such as change in the pitch of the sound. But changes in the phase are still important e.g. locating sound source on the basis of phase difference. This shows that human acoustical mechanism can analyze waveforms directly.

8.2 Cognitive Properties

Upon hearing a certain sound, humans perceive certain information on the basis of physical information but not amplitude or frequency. The extracted information can be general to specific e.g. hearing someone talking or hearing what specifically someone is talking about. This sound comprises of physical information only. But it is still very

difficult to derive information from this physical information such as distinguishing between silence, music, speech or noise from the audio sine wave.

9 Proposed System for Content-Based Rating of Movies

A movie consists of characters, a certain dimension, dialogues and music. We can regard characters as objects, dimension as the subject and dialogues and music as the sound. Here we will focus on the latter features. We will take into account the audio track of the movie to give a certain criterion and ratings based on our algorithm.

As has been mentioned earlier, this paper proposes a method for content-based classification movies. Our methodology is very basic and somewhat limited (partly because there are still a few limitations in ASRs) since our research hasn't matured much. It uses a very simple algorithm to compute the suitable audience for a movie by extracting the audio track of the movie via speech recognition and calculating the percentage of swear words present in the script. The speech-turned-text description is run and compared with the predefined dataset fed to the system. These swear words if they match with the data present in the dataset are recorded. On the basis of this count, the computer performs simple average calculation to derive a percentage of how many words matched the dataset (Fig. 4)..

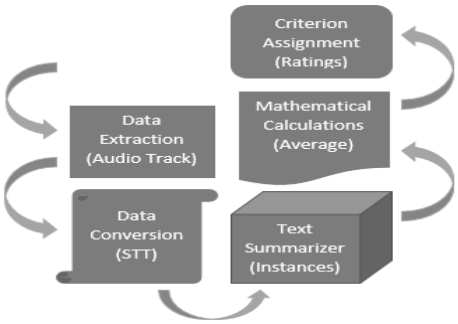

Fig. 4. General pipilien of the system.

Our system shall have the following features:

– Be speaker-independent.
– Be a continuous speech recognizer.
– Supports a text summarizer.
– Be based on a HMM.

All these properties have been carefully complied because a speaker-free system will have a wider range to recognize speech (although it may be only one domain e.g. American movies). With the ability to continuously recognize speech, the system won't face much trouble as the characters speak in a natural manner. As a HMM inspects each

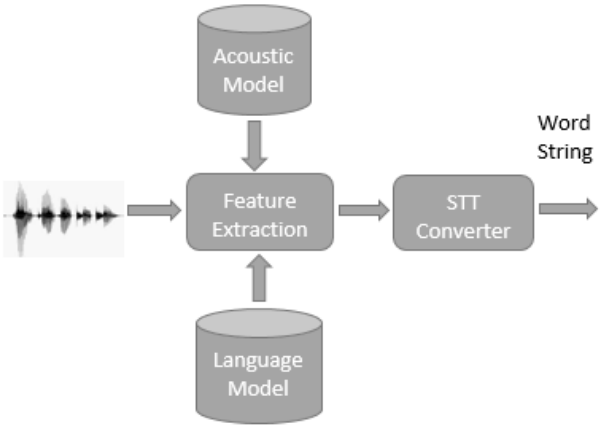

Fig. 5. Signal processing.

Table 1. Features example.

Term	Occurrence	Total
Word 1	07	07
Word 2	05	12
Word 3	09	21
Word 4	11	32
…	…	…
…	…	…
…	…	…

element singularly, therefore it is more reliable as we can assume that the results produced will be more accurate (Fig. 5 and .

Now we move on to how this algorithm actually works. It consists of three steps. Extracting the audio; analyzing and comparing it with the corpus; producing the results and assigning a criterion.

9.1 Audio Track Extraction

First of all, the system extracts the audio track or audio signal of the desired film using speech recognition. The features of the speech recognition system have already been mentioned earlier. After extracting the audio, a textual translation of is prepared via speech to text feature of the system.

9.2 Dataset

The dataset assigned to the system will contain all the information needed to perform the analysis. The problem of large data is resolved as there are only a few abuse or

swear words, may be 500 words in the entire language that are used repeatedly. Therefore the dataset can be made and compiled rather easily. Another interesting feature of the dataset is that it is updateable. If there's a new addition, the dataset can always be updated thereby eliminating any lapses in the analysis.

9.3 Text Analysis of Audio

Once the audio track's text form is achieved, this text form is analyzed by comparing it with the help of a text summarizer by giving in a few key words (in our case swear words) with the dataset that contains all the relevant information. The text summarizer gives all the instances present in the text that match the keyword. The number of occurrences is recorded and stored for further processing.

9.4 Calculating the %age

On the basis of occurrences of swear words recorded in the text, a simple mathematical calculation is performed to compute the %age of swear words present e.g. if in a text of 1000 words, 80 words are swear words, the movie will be given a percentage of 8% on the content scale (Fig. 6).

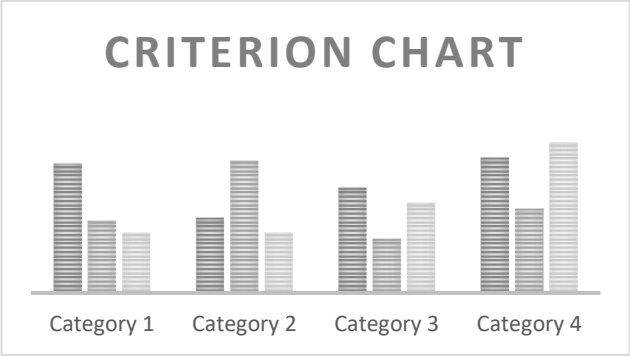

Fig. 6. Classification example.

On the basis of this percentage, the movie shall receive a rating and criterion. Movies can be rated PG-12, PG-13, PG-18+, R-Rated and so on. Each criterion shall be assigned a percentage range e.g. 10%-15% PG-10.

10 Conclusion and Future Work

In this paper, we studied speech recognition, its models, applications, uses and significance in details. We also reviewed audio based classifications and analysis. We came across many proposals ranging from scene segmentation to video frames to audio-visual strategies. At the end, we proposed a theory of our own that indexes online movies according to the percentage of swear words present in the movie's script using speech recognition and other natural language techniques. In the future, we hope to

implement out suggested theory successfully as it can aid a great deal in rating movies on online movies websites such as IMDB and Rotten Tomatoes thereby facilitating the viewers in making a decision and an appropriate choice.

References

1. http://www.lumenvox.com/resources/tips/types-of-speech-recognition.aspx.
2. Lu, L.: Content Analysis for Audio Classification and Segmentation. IEEE Signal Processing Society, 10, pp. 504–516, (2002). doi: 10.1109/TSA.2002.804546.
3. Wold, E., Blum, T., Wheaton, J.: Content-Based Classification, Search and Retrieval of Audio. IEEE Multimedia, 3, pp. 27–36 (1996). doi: 10.1109/93.556537.
4. Pal Singh, P.: Speech Recognition as Emerging Revolutionary Technology. International Journal of Advanced Research in Computer Science and Software Eng., 2(10), pp. 410–413 (2012)
5. Saraceno, C., Leonardi, R.: Audio as a Support to Scene Change Detection and Characterization of Video Sequences, Proc. of ICASSP'97, 4, pp. 2597–2600 (1997)
6. Wang, Y., Huang, J., Liu, Z., Chen, T.: Multimedia Content Classification Using Motion and Audio Information. Proc. of IEEE ISCAS'97, 2, pp. 1488–1491 (1997)
7. Pfeiffer, S., Fischer, S., Effelsberg, W.: Automatic Audio Content Analysis. Proc. ACM Multimedia'96, pp. 21–30 (1996). doi: 10.1145/244130.244139.
8. Nam, J., Tewfik, A.H.: Combined Audio and Visual Streams Analysis for Video Sequence Segmentation. Proc. of ICASSP'97, 3, pp. 2665–2668 (1997)
9. Saunders, J.: Real-Time Discrimination of Broadcast Speech/Music. Proc. of ICASSP'96, 2, pp. 993–996 (1996). doi: 10.1109/ICASSP.1996.543290.
10. Liu, Z., Wang, Y., Chen, T.: Audio Feature Extraction and Analysis for Scene Segmentation and Classification. The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, 20, pp. 61–79 (1998). doi: 10.1023/A:1008066223044.
11. http://en.wikipedia.org/wiki/Hidden_Markov_model.
12. http://en.wikipedia.org/wiki/Dynamic_time_warping.
13. http://en.wikipedia.org/wiki/Artficial_neural_network.
14. http://www.ee.columbia.edu/ln/LabROSA/doc/HTKBook21/node5.html.
15. http://ai-depot.com/ska/paper/node9.html.
16. Mahedero, J.P.G., Martínez, A., Cano, P.: Natural Language Processing of Lyrics. doi: 10.1145/1101149.1101255.
17. Gunsel, B., Tekalp, A.M.: Content-Based Video Abstraction. Department of Electrical Engineering and Center for Electronic Imaging Systems University of Rochester. doi: 10.1109/ICIP.1998.727150.
18. Ferman, A.M., Takalp, A.M., Mehrotra, R.: Effective Content Representation for Video. In: International Conference on Image Processing, 3, pp. 521–525 (1998). doi: 10.1109/ICIP.1998.727251.
19. Tsekeridou, S., Pitas, I.: Audio-Visual Content Analysis for Content-Based Video Indexing. In: IEEE International Conference on Multimedia Computing and Systems, 1, pp. 667–672 (1999). doi: 10.1109/MMCS.1999.779279.
20. Liu, S.C., Bi, J., Jia, Z.Q., Chen, R., Chen, J., Zhou, M.M.: Automatic Audio Classification and Speaker Identification for Video Content Analysis. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007), pp. 91–96 (2007). doi: 10.1109/SNPD.2007.516.

21. Huang, C., Fu, T., Chen, H.: Text-Based Video Content Classification for Online Video-Sharing Sites. Journal of the American Society for Information Science and Technology, 61(5), pp. 891–906. doi: 10.1002/asi.21291.
22. Huang, J., Kingsbury, B., Mangu, L., Padmanabhan, M., Saon, G., Zwei, G.: Recent Improvements In Speech Recognition Performance On Large Vocabulary Conversational Speech (Voicemail And Switchboard). In: Sixth International Conference on Spoken Language Processing (2000). doi: 10.21437/ICSLP.2000-819.
23. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.C.: Short-Term Audio-Visual Atoms for Generic Video Concept Classification. Electrical Engineering Department
24. Nakamura, Y., Kanade, T.: Semantic Analysis for Video Contents Extraction - Spotting by Association in News Video. In: Conference: Proceedings of the Fifth ACM International Conference on Multimedia '97 (1997). doi: 10.1145/266180.266391.
25. White, D.J., King, A.P., Duncan, S.D.: Voice Recognition Technology as a Tool for Behavioral Research. Behav. Res. Methods Instrum. Comput., 34(1), pp. 1–5 (2002). doi: 10.3758/bf03195418.

ISSN 1870-4069

Wildfire Risk Assessment through Machine Learning-Based Metamodels

Pedro Adrián Ibarra-Elizondo, Susana Favela-Lara

Universidad Autónoma de Nuevo León, Facultad de Ciencias Biológicas,
San Nicolás de los Garza, NL,
Mexico

pedro.ibarraeo@uanl.edu.mx, susana.favelalr@uanl.edu.mx

Abstract. One of the main threats to tropical ecosystems is the occurrence of wildfires, which simultaneously endanger animal communities and human health due to the destructive nature of the phenomenon. To address this, a variety of disaster prediction algorithms have been developed; however, these still present gaps that limit their precision and applicability. In response, this study proposes to take advantage of the physical and biological complexity of the "Cumbres de Monterrey" National Park, located in Nuevo León, Mexico, for the categorization of its territory based on a fire risk assessment using ensemble metamodels that integrate seven Machine Learning algorithms. These models are trained with wildfire records from the fire seasons between 2013 and 2022, along with the spatiotemporal variation of physical, biological, and demographic components. The predictive potential of both models and metamodels is evaluated using six different metrics. The comparison reveals that the metamodels exhibit fluctuating effectiveness in classifying risk zones in the region, ranging from 73.77% in 2020 to 91.3% in 2019. The results provide a robust methodological framework for wildfire prediction, overcoming the limitations of traditional Machine Learning approaches through the implementation of ensemble-based metamodels.

Keywords: Artificial Intelligence, forest fires, conservation, remote sensing, wildfire prediction.

1 Introduction

Fire has remained a key factor in the development of human societies and in the evolution of ecosystems worldwide, due to its influence on ecological processes such as species regeneration, nutrient cycling, and landscape configuration (Bowman *et al*., 2009; Moritz *et al*., 2014).

However, when this natural and stochastic phenomenon escapes human control, it is referred to as a wildfire. In particular, those affecting areas with natural vegetation cover are classified as forest fires, which generate significant disturbances in both ecological and anthropogenic systems. One such victim of these incidents is the Cumbres de Monterrey National Park (CMNP), located in the state of Nuevo León,

México, has been affected by these incidents (Fueyo-MacDonald, 2013; CONANP, 2020). In most cases, the origins of these disasters are anthropogenic, such as waste pollution from the Monterrey Metropolitan Area (Fueyo-MacDonald, 2013), soil stress, and the repurposing of areas for tourism activities (Narváez-Torres & Lazcano-Villareal, 2013).

Furthermore, due to its highly rugged geomorphology, this region exhibits a high degree of climatic dynamism (Alanís-Flores & Velazco-Macías, 2013), which causes variability in average temperatures and annual precipitation levels that influences the persistence of wildfire events (Jiménez-Pérez *et al*., 2013).

The highly dynamic and destructive nature of forest fires makes it imperative to implement preventive measures to avoid their uncontrolled spread (McKenzie *et al*., 2004; Moritz *et al*., 2014). Currently, there are regional action protocols in place for the immediate management of these events, such as the Fire Management Program (SCMF, 2017). However, their effectiveness critically depends on the availability of reliable predictive models that allow for the optimal distribution of primary response resources (Stocks & Martell, 2016; McGovern *et al*., 2017; Tymstra *et al*., 2020).

Over the last two decades, with advances in data analysis and artificial intelligence, multiple methodologies have been proposed to achieve reliable predictions of wildfires in natural areas. These range from simple models that incorporate and synthesize expert perspectives such as Multi-criteria Decision Making (MCDM) (Eskarandi *et al*., 2015); basic statistical archetypes like Logistic or Multinomial Regression (Thai Pham *et al*., 2020); to more complex algorithms that include Machine Learning methods such as Neural Networks (Barmpoutis et al., 2020), Support Vector Machines (SVM) (Özbayoglu & Bozer, 2012), Fuzzy Logic (Neuro-fuzzy) (Thai Pham *et al*., 2020), Bayesian Networks (Thai Pham *et al*., 2020), and Random Forests (Arkin *et al*., 2019), among others (Abid, 2021; Preeti *et al*., 2021).

The main challenge in establishing these models lies in the strong dependency of wildfires on climatic variables (McKenzie et al., 2004; Moritz *et al*., 2014), which introduces a substantial uncertainty component in occurrence modeling. For this reason, these models frequently analyze the physical factors of the terrain (climatology, topography, and hydrology) (Jain *et al*., 2020; Malik *et al*., 2021; Xie et al., 2022). However, only a few models consider biological factors (fuel type and load, vegetation composition, land use) (Cencerrado *et al*., 2014; Lauer *et al*., 2017), while demographic factors are often actively overlooked (Kondylatos *et al*., 2022).

These omissions create significant predictive gaps, as they are unable to objectively analyze the multidimensional complexity of forest fires (Cheng & Wang, 2008; Kozik *et al*., 2013; Jain *et al*., 2020). As a result, it becomes necessary to synergistically integrate a considerable number of base models whose predictive strengths mutually compensate for their individual limitations. This can be achieved through the development of ensemble metamodels, which combine heterogeneous architectures (physical, statistical, and machine learning models) for the development and selection of multiple submodels through systematic variation of hyperparameters (tuning) and consensus algorithms. These models also implement adaptive weighting mechanisms that optimize the relative contribution of each submodel, generating consolidated estimates with lower bias and variance than individual models (Nguyen *et al*., 2021).

Currently, there are few initiatives in the region exploring the application of autonomous machine learning algorithms for forest fire prediction. This scarcity has led to uncertainty regarding their operational viability and actual effectiveness among the stakeholders responsible for the conservation of the CMNP.

Given this context, the present study proposes to develop an integrative metamodel that generates predictive consensus from multiple autonomous algorithms, quantitatively evaluates the classification capacity of vulnerable areas using spatially explicit metrics, estimates risks based on hazard and exposure indicators, and contributes to the technical debate on the applicability of these methodologies in protected natural systems.

2 Methodology

2.1 Study Area

The Cumbres de Monterrey region, designated as a Protected Natural Area (ANP, for its initials in Spanish) with National Park status on November 24, 1939 (DOF, 1939), currently encompasses an area of 1,779.95 km² and is located across the states of Nuevo León and Coahuila de Zaragoza (Fig. 1). This designation was later redefined by an ordinance issued on November 17, 2000 (DOF, 2000).

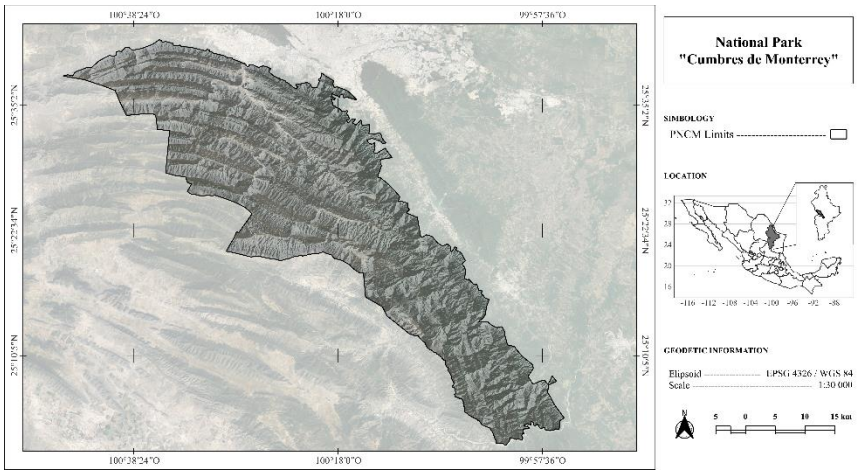

Fig. 1. Geographic location of the ANP "Cumbres de Monterrey" National Park.

2.2 Fire Records

Through a query conducted on the databases of the Fire Danger Prediction System (SPPIF, for its initials in Spanish) (CONAFOR, 2024) and the Fire Information for Resource Management System (FIRMS) (NASA, 2024), historical records of wildfires occurring within the CMNP during the fire season -February to May- from 2013 to 2022 were collected.

A similar amount of pseudo-absence points *(background points*) was artificially added to the true wildfire records for each year, ensuring spatial randomness under the condition that their origin would not fall outside the delimited region nor overlap geographically with the actual records.

2.3 Predictors

Climatic. To measure regional temperature and humidity factors, annual aerial imagery from the Landsat 8 OLI/TIRS satellite was used, restricting the selection to photographs produced between February and May, considering only those with minimal cloud cover.

The analysis of the Land Surface Temperature (LST) is applied (Eq. 1):

$$LST = \frac{T_B}{1+(\lambda \cdot \frac{T_B}{\rho \cdot \ln \epsilon})} \tag{1}$$

$\mathbf{T_B}$ = Brightness temperature or radiance
$\boldsymbol{\lambda}$ = Wavelength of emitted radiance (11.5 μm)
$\boldsymbol{\rho}$ = $h \times c/\sigma = 1.438 \times 10^{-2}$ mK^1
$\boldsymbol{\varepsilon}$ = Surface emissivity

Similarly, for the estimation of humidity in the National Park, the analysis of the Normalized Difference Moisture Index (NDMI) is used (Eq. 2):

$$NDMI = \frac{(NIR-SWIR)}{(NIR+SWIR)} \xrightarrow{In\ Landsat\ 8} \frac{(Band\ 5-Band\ 6)}{(Band\ 5+Band\ 6)} \tag{2}$$

Topographic. The variance in elevation of the topography is measured using the N25W100 and N25W101 digital elevation models extracted by the Shuttle Radar Topography Mission (SRTM) (NASA, 2000), through which the analysis of slope inclinations and aspect is performed, considering solar incidence.

Hydrological. Each of the surface watercourses within the CMNP is examined and georeferenced to estimate the Euclidean distance (proximity) from each point in the region to these watercourses.

Ecosystem-based. Using satellite imagery from the Landsat 8 OLI/TIRS satellite, annual Normalized Difference Vegetation Indices (NDVI) relevant to the same time periods established in this study are obtained (Eq. 3):

$$NDVI = \frac{(Red-NIR)}{(Red+NIR)} \xrightarrow{In\ Landsat\ 8} \frac{(Band\ 5-Band\ 4)}{(Band\ 5+Band\ 4)} \tag{3}$$

σ = Boltzmann constant, h = Planck constant, c = Speed of light.

Additionally, the Land Use and Vegetation Classification Series VII (INEGI, 2021) was employed to categorize the main ecosystems and urban areas within the CMNP territory.

Demographic. The proximity of each territorial point to the localities listed in the 2020 Population and Housing Census (INEGI, 2021) and to tourist routes recorded in the AllTrails platform up to 2022 was analyzed.

2.4 Base Models

To construct the metamodel, the following base ensemble models were defined and trained in parallel using the R v.4.4.3 "Trophy Case" environment (Table 1):

Table 2. Structure for defining the base ensemble models.

Model	Permutation Levels	Library	Function	Arguments	Total Submodels
RL	3	glmnet	logistic_reg	— penalty — mixture	9
AD	3	rpart	decision_tree[2]	— cost_complexity — tree_depth — min_n	18
SVM	3	kernlab	svm_rbf	— cost — rbf_sigma	9
KNN	3	kknn	nearest_neighbor	— neighbors — weight_func — dist_power	27
RM	3	glmnet	multinom_reg	— penalty — mixture	9
BA	3	ranger	rand_forest[3]	— min_n — trees	6
XGB	3	xgboost	boost_tree[3]	— learn_rate — trees — tree_depth	18

Standard Logistic Regression (LR). This model is based on adjusting the β coefficients to minimize the difference between estimated probabilities and the actual observed values (Eq. 4):

$$P(Y = 1|X) = \frac{1}{1+e^{(-\beta_0+\beta_1 x_{i1} \cdots \beta_n x_{in})}} \quad (4)$$

$\boldsymbol{\beta_n}$ = Coefficient of parameter *n*.
$\mathbf{x_n}$ = Value of predictor *n*.

[2] Only archetypes with *tree_depth* > 1 are selected.
[3] Only archetypes with *trees* > 1 are selected.

Decision Tree (DT). These are supervised learning algorithms whose decision-making process is sequential -that is, it follows a specific order and hierarchy- and considers the characteristics of the input data for classification.

The process starts at a root node, which applies the first split condition to the dataset. Each subsequent decision node partitions the data based on specific attributes, and the branches represent the possible outcomes of each condition. It ends in leaf nodes, which contain the final predictions: discrete classes for classification problems or continuous values for regression.

For the establishment of the tree, in the case of classification issues specifically, metrics based on entropy are used (Eq. 5):

$$H(X) = \sum_{i=1}^{n} p_i \log_2(p_i) \quad (5)$$

H(X) = Entropy function of dataset X.
p_i = Proportion of examples of class *i* in the set X ($0 \leq p_i \leq 1$).

Support Vector Machines (SVM). Define a hyperplane in an n-dimensional space (being a line in a two-dimensional space or a plane in a three-dimensional space) that maximizes the margin between two distinct categories, that is, the distance between the hyperplane and the closest data samples –the fire records– from each class, defined as the support vectors (Eq. 6):

$$maximize \frac{2}{\|w\|} subject\ to\ y_i(w^T x_i + b) \geq 1 \quad (6)$$

w = Weight vector
T = Transposition of the vector or matrix.
x_i = Feature vector of observation *i*.
y_i = Label or category of observation *i* ($y_i \in \{-1,1\}$).
b = Bias term

The final classification is based on the calculation of the observation's distance to the hyperplane, specifically on the sign (positive or negative) of that distance.

k-Nearest Neighbors (KNN). This algorithm classifies an observation based on the training data that are closest (nearest neighbors) to it within an n-dimensional feature space. A parameter $k \in \{\mathbb{Z} > 0\}$ can be defined to determine the number of neighbors considered in the final decision-making process. It relies on distance metrics to calculate the closeness between points, with a common example being the Euclidean distance (Eq. 7):

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (7)$$

x_i, x_j = Feature vectors of observations *i* and *j*.
n = Number of features or predictors.

Multinomial Regression (MR). This is a direct extension of linear regression that allows modeling nonlinear relationships between dependent and independent variables by introducing polynomial terms. It fits a curve to a specific dataset.

A polynomial is a mathematical expression that sums terms in which an independent variable x is raised to various powers, also referred to as degrees. In general, a polynomial of degree *n* is expressed as (Eq. 8):

$$y = \beta_0 + \beta_1 x_{i1} + \beta_1 {x_{i1}}^2 + \beta_1 {x_{i1}}^3 \ldots \beta_n {x_{in}}^n \quad (8)$$

$\boldsymbol{\beta_n}$ = Coefficient of parameter *n*.
$\mathbf{x_n}$ = Value of predictor *n*.

The resulting model maintains linearity with respect to the parameters β, allowing the algorithm to apply the same techniques used in ordinary linear regression to fit the prediction model.

Random Forest (RF). Its structure is based on multiple decision trees -hence the name forests- which are trained to generate individual predictions that are then aggregated to produce a final overall prediction.

One of the core principles for achieving robust modeling is the introduction of randomness to reduce bias, typically drawn from two main sources: a Bootstrap Aggregating (Bagging) algorithm or a random selection of features. Similar to decision trees, Random Forest models are built using splitting criteria and entropy-based measures (see Eq. 5).

Extreme Gradient Boosting (XGBoost). An advanced implementation of gradient boosting that combines multiple weak learners to sequentially build a stronger model. It is specifically optimized for flexibility, efficiency, and accuracy.

It is trained sequentially using decision trees. In each iteration, the model is adjusted to predict the errors made by the previous model (Eq. 9):

$$F_{m+1}(x) = F_m(x) + \gamma \cdot h(x) \quad (9)$$

$\mathbf{F_m(x)}$ = Model at iteration *m*.
$\mathbf{h(x)}$ = Tree fitted at the current iteration to correct the residuals of $F_m(x)$.
$\boldsymbol{\gamma}$ = Learning rate ($0 \leq \gamma \leq 5$).

A *logarithmic loss function* is optimized using gradient descent (Eq. 10):

$$L(y, \hat{y}) = \sum_{i=1}^{n} \ell\,(y_i, \hat{y}_i) + \sum_k \Omega\,(f_k) \quad (10)$$

$\mathbf{L(y)}$ = Total objective function.
$\boldsymbol{\ell(y_i)}$ = Loss function for observation *i*.
$\boldsymbol{\Omega(f_k)}$ = Regularization term for tree f_k (see Eq. 11).

This regularization term prevents overfitting on the training data, as defined (Eq. 11):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (11)$$

$\boldsymbol{\gamma}$ = Penalty for the number of leaves *T* in the tree ($\gamma \geq 0$).

$\mathbf{w_j}$ = Weights of the leaves.
$\boldsymbol{\lambda}$ = Regularization coefficient that penalizes the magnitude of w_j ($\lambda \geq 0$).

2.5 Training and Evaluation of Submodels

The constructed datasets are split into 70% for training the base models and 30% for testing. Additionally, a *k-fold* cross-validation algorithm (k = 10) is applied to the training subset in order to select the best-performing submodels based on various statistical metrics (see Table 2).

Table 2. Statistical Metrics for Evaluation and Selection of the Best Submodels.

Statistic	Formula[4]
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2 \cdot Precision \cdot Sensitivity}{Precision+Sensitivity}$
Kappa	$\frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP+FP) \cdot (FP+TN) + (TP+FN) \cdot (FN+TN)}$
Matthews Correlation Coefficient	$\frac{(VP \cdot VN - FP \cdot FN)}{\sqrt{(VP+FP) \cdot (FP+VN) \cdot (VP+FN) \cdot (FN+VN)}}$

2.6 Feature Importance

Gain. The individual contributions of each predictor to each model are evaluated through a feature importance permutation metric. This technique estimates the importance (gain) of each variable in a trained model by calculating the change in model performance when predictor values are randomly assigned.

2.7 Construction and Application of the Metamodel

The prediction results are generalized through a logistic regression metamodel (Eq. 12-13):

$$p_i = \frac{1}{1+e^{(-z_i)}} \tag{12}$$

$\mathbf{p_i}$ = Probability that an event belongs to a category, in this case, a fire-prone area ($0 \leq p_i \leq 1$).
$\mathbf{z_i}$ = Linear combination of the predictors (see Eq. 14).

[4] *TP = True Positives*; *TN = True Negatives*; *FP = False Positives*; *FN = False Negatives*.

$$z_i = \beta_0 + \beta_1 x_{i1} \ldots \beta_n x_{in} \tag{13}$$

$\boldsymbol{\beta_n}$ = Coefficient of parameter *n*.
$\mathbf{x_n}$ = Value of predictor *n*.

This algorithm is fitted by maximizing the likelihood, selecting the coefficients (β) that provide the best fit for the observed data (Eq. 14):

$$\ln L(\beta) = \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i)] \tag{14}$$

L(β) = Likelihood function.
$\mathbf{y_i}$ = Dependent variable or label of observation *i* ($y_i \in \{0, 1\}$).

The addition of the Least Absolute Shrinkage and Selection Operator (*Lasso*) technique is considered, enforcing the sum of the absolute values of the coefficients (Eq. 15):

$$\lambda \sum_{j=1}^{p} |\beta_j| \tag{15}$$

λ = Regularization/penalty parameter.

Therefore, the process is given by (Eq. 16):

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{-[\sum_{i=1}^{n} (y_i \ln(p_i) + (1 - y_i) \cdot \ln(1 - p_i))] + \lambda \sum_{j=1}^{p} |\beta_j|\} \tag{16}$$

$\hat{\boldsymbol{\beta}}$ = Vector of estimated coefficients.
argmin = Argument of the minimum coefficients.

This process regularizes and selects the variables more effectively, minimizing the risk of overfitting and improving the model's interpretability.

Finally, the logistic regression algorithm enables the classification of fire-prone areas based on the geospatial variance of the same predictors considered in the base models (Fig. 2).

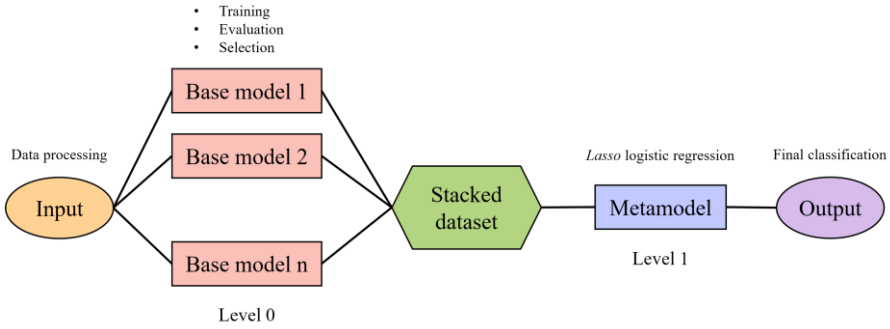

Fig. 2. Construction Process of Base Algorithms and the Metamodel for Classification.

3 Results

3.1 Fire Incident Records and Databases

Considering the selected time range from 2013 to 2022, a total of 4,694 confirmed fire records were identified between the months of February and May, varying in intensity, within the CMNP region (Fig. 3).

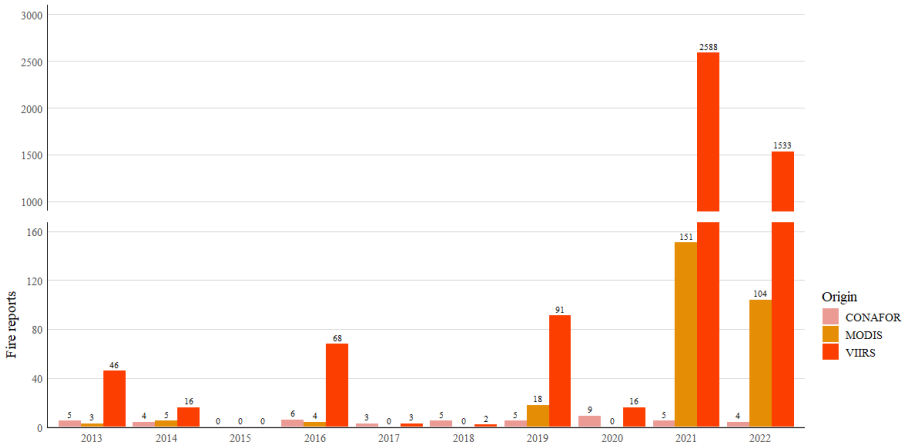

Fig. 3. Fire reports within CMNP from 2013 to 2022.

The analysis for the years 2015, 2017, and 2018 was excluded due to the insufficient availability of data to establish any modeling archetype ($n < 10$), thus rendering the modeling process impractical. From the 4,694 records collected from the SPPIF and FIRMS catalogs, seven databases corresponding to the years considered in the study were generated (Table 3).

Table 3. Data arrangement for each database is constructed annually.

Year	Occurrences	Pseudo-absences	Total
2013	54	54	108
2014	35	35	50
2016	78	78	156
2019	114	114	228
2020	25	25	50
2021	2 744	2 744	5 488
2022	1 641	1 641	3 282

3.2 Base Models and Submodels

Machine learning algorithms demonstrated their effectiveness in classification tasks, as evidenced by the performance evaluation metrics (Fig. 4).

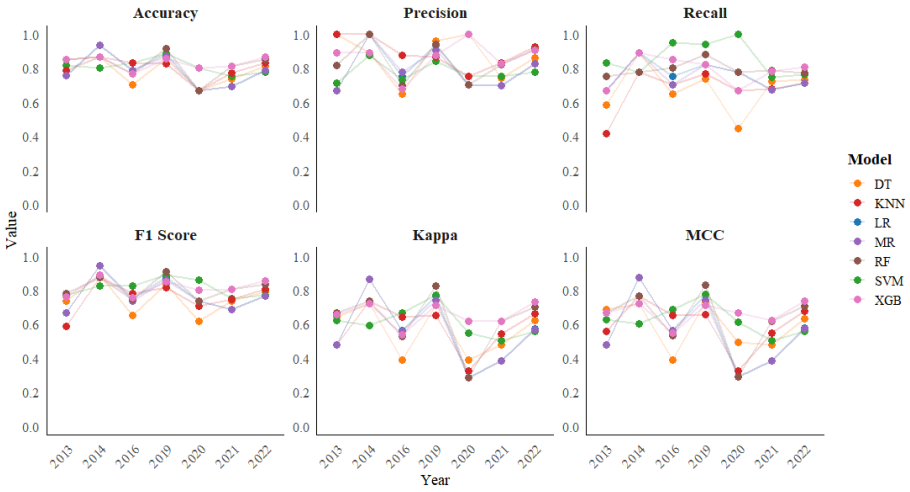

Fig. 4. Temporal dynamics of model performance.

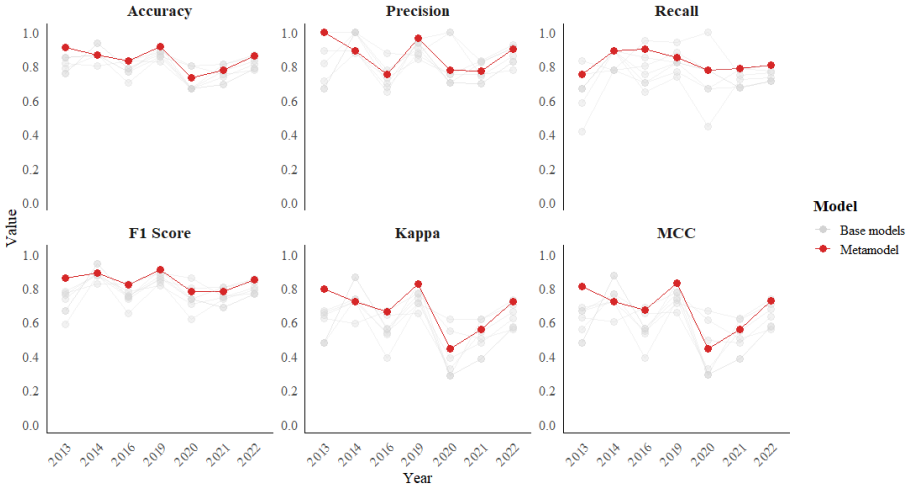

Fig. 6. Evaluation results of the constructed metamodels.

3.3 Predictor Importance

Regarding the relevance of predictors for the classification potential of the trained base models -and consequently the metamodels- the results of the permutation importance method applied for each year are presented (Fig. 5).

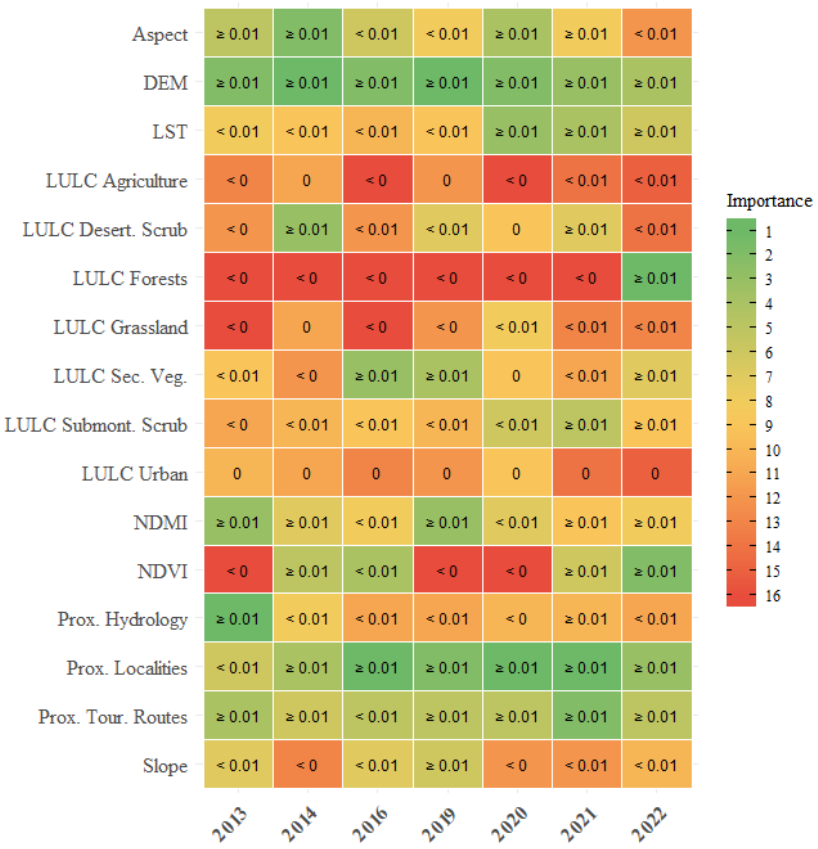

Fig. 5. Temporal dynamics of predictor importance used in the model structures. Where: DEM = Digital Elevation Model; LST = Land Surface Temperature; LULC = Land Use Land Cover; Sec. = Secondary Vegetation; NDMI = Normalized Difference Moisture Index; NDVI = Normalized Difference Vegetation Index.

3.4 Metamodels

Coefficients. Comparative analysis of the models (Table 4) shows a recurring pattern in which, during some stages, the influence of certain algorithms (such as Logistic Regression, Multinomial Regression, and Random Forests) is minimized or nullified (coefficients equal to 0), while others (Support Vector Machines, Extreme Gradient Boosting, and k-Nearest Neighbors) are favored with higher coefficients.

However, in specific periods, the metamodel incorporates all base algorithms, albeit with unequal weightings. Overall, there is a trend of preference for advanced techniques (such as XGB and SVM) over classical models (Logistic/Multinomial Regression), adjusting their influence according to the evaluated design.

Table 4. Determination of coefficients for the establishment of the metamodels.

	Coefficient							
Year	Intercept	RL	AD	SVM	KNN	RM	BA	XGB
2013	-1.05	0.00	0.36	0.95	0.18	0.00	0.18	0.44
2014	-1.50	0.00	0.31	0.81	0.30	0.69	0.01	0.94
2016	-2.02	0.00	0.65	1.48	0.63	0.00	0.00	0.76
2019	-2.38	0.24	0.72	0.93	0.55	0.49	0.64	1.08
2020	-2.25	0.49	0.08	1.34	0.83	0.89	0.00	0.50
2021	-1.55	0.00	0.19	0.54	0.64	0.00	0.55	0.94
2022	-2.04	0.14	0.34	0.53	0.64	0.15	0.59	1.08

Evaluation. Over the years, the metamodel has shown fluctuating performance relative to the base algorithms, reflecting both advancements and limitations (Fig. 6). Although the metamodel has evolved toward greater classification efficiency, its performance remains relative, depending on the integration capability of the base algorithms and its direct competition with more powerful methods.

3.5 Spatial evaluation

The categorized susceptible territory shows a spatially homogeneous distribution in the region, suggesting that wildfire occurrence probability may largely be attributed to stochastic events. However, starting in 2016, clusters of high susceptibility become evident in the central area of the region (Fig. 7).

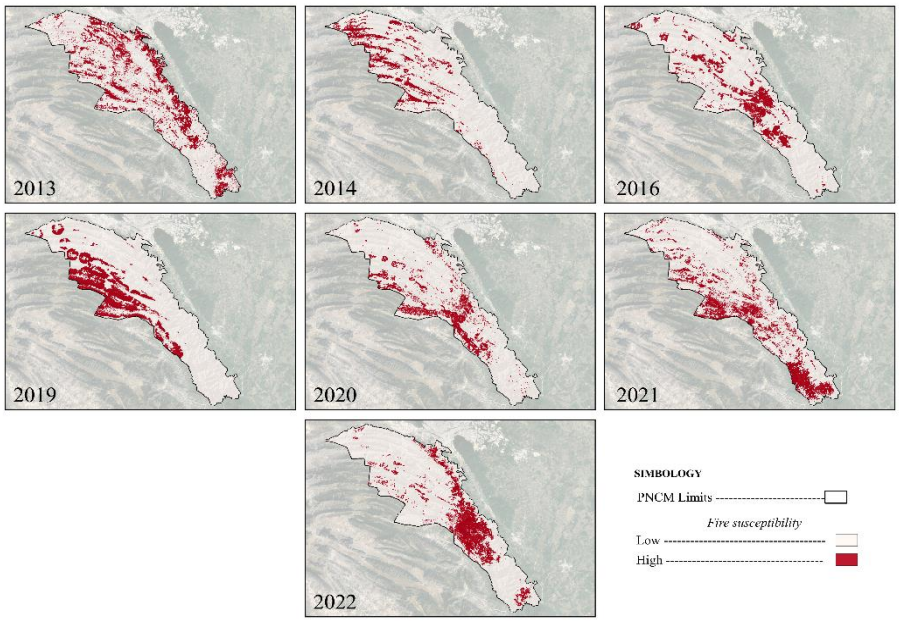

Fig. 7. Wildfire susceptibility within the CMNP.

Similarly, it is possible to identify specific territorial areas susceptible to wildfires during the different established periods (Table 5).

Table 4. Extent of wildfire-susceptible areas within the CMNP during each period.

Year	Susceptible area ($km^{2)}$
2013	422.42
2014	228.57
2016	261.19
2019	354.04
2020	252.47
2021	387.47
2022	320.66

4 Discussion

4.1 Methodological Innovation

This research addresses the methodological challenge of non-integrative application of machine learning models, where the individual deficiencies of each model type are often highlighted (Nguyen *et al*., 2021; Kondylatos *et al*., 2022; Xie *et al*., 2022). This is achieved by constructing metamodels that integrate multiple Machine Learning algorithms to evaluate and predict the occurrence of wildfires for each year. These metamodels simultaneously incorporate LR, DT, SVM, KNN, MR, RF, and XGBoost algorithms, allowing for the creation of highly specialized classification models.

To logically classify regions with high susceptibility to wildfire occurrence, it is essential to rigorously select the predictors used to train the model (Reichstein *et al*., 2019; Jain *et al*., 2020). In most cases, this selection is empirical, which minimizes the possibilities of a correct analysis of the factors influencing the occurrence of wildfires, as it does not allow the evaluation of new and different predictors (Jain *et al*., 2020). For example, few studies incorporate demographic predictors (Kondylatos *et al*., 2022), which overlooks the anthropogenic impact on fire occurrence in natural regions, thus introducing a considerable bias when defining effective proposals for fire prevention and control. It is impossible to separate the human attributes that interact with the landscape (Likens, 1991; Western, 2001). A secondary finding of this study, resulting from the inclusion of such predictors in the training of models, is the identification of the temporal dynamics of the importance of specific predictors over the years.

4.2 Precision of Results

The developed metamodels allow for precise categorization of wildfire risk in the "Cumbres de Monterrey" National Park, with predictive accuracy varying across the years, reaching 91.30% in its best performance in 2019. These results enable the identification of high-vulnerability areas, promoting the implementation of appropriate prevention and response strategies.

However, it is necessary to mention that the model's accuracy varies over time, with reductions in 2020 and 2021 (73.77% and 77.90%, respectively). These variations could reflect changes in the quality and quantity of available data, such as the scarcity of fire records in certain years (e.g., 2015, 2017, and 2018).

This suggests the need for continuous improvement in the quality of training data and the tuning of model hyperparameters (Preeti *et al*., 2021), as these factors affect the models' ability to generate robust predictions depending on their inherent complexity (Brigato & Iocchi, 2021). Nevertheless, advanced techniques such as *k-fold* cross-validation and permutation-based feature selection ensure that the models are appropriately adjusted to the fire dynamics of the region (Tzu-Tsung & Po-Yang, 2020).

5 Conclusion

This study demonstrates the ability to characterize the territory of the CMNP through wildfire risk assessment using autonomous machine learning metamodels. The study underscores the need to coherently define temporal-spatial dimensions for any study based on its objectives, as this ensures the correct collection of predictor information for model training. Secondly, the integration of sufficient classification models -and iterations of these models- will increase classification accuracy. Lastly, rigorous analysis schemes must be developed for the interpretation of results without subjectivity.

The trajectory suggests that, while no single algorithm is universally superior, gradient-based and ensemble methods have the potential to redefine classification excellence standards. The methodology in this study can establish a technological innovation in addressing the wildfire problem in Mexico's natural regions, while simultaneously promoting the use of Geographic Information Systems (GIS), real-time spatial analysis, and autonomous Machine Learning methods for strategic planning in the control of environmental issues.

6 Declaration of interests

The author declares no financial interests or known relationships that may have influenced the work presented in this article.

Acknowledgments

This work was supported by the National Council of Humanities, Sciences, and Technologies of Mexico (SECIHTI) with support number 4005985.

Acknowledgments are extended to the Department of Ecology at the Universidad Autónoma de Nuevo León, to biologists Adrián González Martínez, Nora Guadalupe Niño Olguín, and Claudia Cecilia Vargas Torres, and to translator Valeria Azeneth Balderrama Sauceda.

References

1. Abid, F. (2021). A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems. Fire Technology, 57(2), 559–590. https://doi.org/10.1007/s10694-020-01056-z
2. Alanís-Flores, G., & Velazco-Macías, C. (2013). Tipos de Vegetación. En Historia Natural del Parque Nacional Cumbres de Monterrey, México (pp. 207–220). UANL - CONANP.
3. Arkin, J., Coops, N. C., Hermosilla, T., Daniels, L. D., & Plowright, A. (2019). Integrated fire severity–land cover mapping using very-high-spatial-resolution aerial imagery and point clouds. International Journal of Wildland Fire, 28(11), 840. https://doi.org/10.1071/WF19008
4. Barmpoutis, P., Stathaki, T., Dimitropoulos, K., & Grammalidis, N. (2020). Early Fire Detection Based on Aerial 360-Degree Sensors, Deep Convolution Neural Networks and Exploitation of Fire Dynamic Textures. Remote Sensing, 12(19), 3177. https://doi.org/10.3390/rs12193177
5. Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C., Roos, C. I., Scott, A. C., … Pyne, S. J. (2009). Fire in the Earth System. Science, 324(5926), 481–484. https://doi.org/10.1126/science.1163886
6. Brigato, L., & Iocchi, L. (2021). A Close Look at Deep Learning with Small Data. 2020 25th International Conference on Pattern Recognition (ICPR), 2490–2497. https://doi.org/10.1109/ICPR48806.2021.9412492
7. Cencerrado, A., Cortés, A., & Margalef, T. (2014). Response time assessment in forest fire spread simulation: An integrated methodology for efficient exploitation of available prediction time. Environmental Modelling & Software, 54, 153–164. https://doi.org/10.1016/j.envsoft.2014.01.008
8. Cheng, T., & Wang, J. (2008). Integrated Spatio-temporal Data Mining for Forest Fire Prediction. Transactions in GIS, 12(5), 591–611. https://doi.org/10.1111/j.1467-9671.2008.01117.x
9. CONAFOR. (2024). Sistema de Predicción de Peligro de Incendios Forestales (SPPIF). [Database]. SPPIF.
10. CONANP. (2020). Programa de Manejo. Parque Nacional Cumbres de Monterrey. CONANP. https://www.conanp.gob.mx/anp/consulta/Borrador%20PM%20PN%20Cumbres%20Mty%20para%20Consulta%20Pública%202020.pdf
11. DOF. (1939). Decreto que declara Parque Nacional "Cumbres de Monterrey", los terrenos que rodean a dicha población. DOF. https://www.dof.gob.mx/nota_to_imagen_fs.php?cod_diario=191064&pagina=10&seccion=1
12. DOF. (2000). Decreto por el que se declara área natural protegida, con el carácter de parque nacional, la región conocida con el nombre de Cumbres de Monterrey, ubicada en los municipios de Allende, García, Montemorelos, Monterrey, Rayones, Santa Catarina, Santiago y San Pedro Garza García, Estado de Nuevo León. DOF. https://dof.gob.mx/nota_detalle.php?codigo=2063788&fecha=17/11/2000#gsc.tab=0

13. Eskarandi, S., Oladi-Ghadikolaei, J., Jalilvand, H., & Reza-Saradjian, M. (2015). Prediction of Future Forest Fires using the MCDM Method. Polish Journal of Environmental Studies, 24(5), 2309–2314.
14. Fueyo-MacDonald, L. (2013). Las Áreas Naturales Protegidas en México y el Parque Nacional Cumbres de Monterrey. En Historia Natural del Parque Nacional Cumbres de Monterrey, México (pp. 37–40). UANL - CONANP.
15. INEGI. (2021). Marco Geoestadístico, diciembre 2021. [Database]. Marco Geoestadístico. https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463849568
16. Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020a). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019
17. Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020b). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019
18. Jiménez-Pérez, J., Aguirre-Calderón, O., Yerena-Yamallel, I., & Alanís-Rodríguez, E. (2013). Cambio Climático. En Historia Natural del Parque Nacional Cumbres de Monterrey, México (pp. 207–220). UANL - CONANP.
19. Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-Torres, M., & Carvalhais, N. (2022). Wildfire Danger Prediction and Understanding with Deep Learning. Geophysical Research Letters, 49(17), e2022GL099368. https://doi.org/10.1029/2022GL099368
20. Kozik, V. I., Nezhevenko, E. S., & Feoktistov, A. S. (2013). Adaptive prediction of forest fire behavior on the basis of recurrent neural networks. Optoelectronics, Instrumentation and Data Processing, 49(3), 250–259. https://doi.org/10.3103/S8756699013030060
21. Lauer, C. J., Montgomery, C. A., & Dietterich, T. G. (2017). Spatial interactions and optimal forest management on a fire-threatened landscape. Forest Policy and Economics, 83, 107–120. https://doi.org/10.1016/j.forpol.2017.07.006
22. Likens, G. E. (1991). Some Consequences of Long-Term Human Impacts on Ecosystems. Revista Chilena de Historia Natural, 64, 597–614.
23. Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., Chiao, S., & Gao, J. (2021). Data-Driven Wildfire Risk Prediction in Northern California. Atmosphere, 12(1), 109. https://doi.org/10.3390/atmos12010109
24. McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., & Williams, J. K. (2017). Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. Bulletin of the American Meteorological Society, 98(10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1
25. McKENZIE, D., Gedalof, Z., Peterson, D. L., & Mote, P. (2004). Climatic Change, Wildfire, and Conservation. Conservation Biology, 18(4), 890–902. https://doi.org/10.1111/j.1523-1739.2004.00492.x
26. Moritz, M. A., Batllori, E., Bradstock, R. A., Gill, A. M., Handmer, J., Hessburg, P. F., Leonard, J., McCaffrey, S., Odion, D. C., Schoennagel, T., & Syphard, A. D. (2014). Learning to coexist with wildfire. Nature, 515(7525), 58–66. https://doi.org/10.1038/nature13946

27. Narváez-Torres, S., & Lazcano-Villareal, D. (2013). Anfibios y Reptiles. En Historia Natural del Parque Nacional Cumbres de Monterrey, México (pp. 207–220). UANL - CONANP.
28. NASA. (2000). Shuttle Radar Topography Mission (SRTM). [Database]. SRTM.
29. NASA. (2024). NASA's Fire Information for Resource Management System (FIRMS). [Database]. FIRMS.
30. Nguyen, P. T., Di Rocco, J., Iovino, L., Di Ruscio, D., & Pierantonio, A. (2021). Evaluation of a machine learning classifier for metamodels. Software and Systems Modeling, 20(6), 1797–1821. https://doi.org/10.1007/s10270-021-00913-x
31. Özbayoğlu, A. M., & Bozer, R. (2012). Estimation of the Burned Area in Forest Fires Using Computational Intelligence Techniques. Procedia Computer Science, 12, 282–287. https://doi.org/10.1016/j.procs.2012.09.070
32. Pham, B. T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H. P. H., Phong, T. V., Nguyen, D. H., Le, H. V., Mafi-Gholami, D., Prakash, I., Thi Thuy, H., & Tuyen, T. T. (2020). Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction. Symmetry, 12(6), 1022. https://doi.org/10.3390/sym12061022
33. Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., & Sudi, A. (2021). Forest Fire Prediction Using Machine Learning Techniques. 2021 International Conference on Intelligent Technologies (CONIT), 1–6. https://doi.org/10.1109/CONIT51480.2021.9498448
34. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1
35. SCMF. (2017). Programa de Manejo del Fuego. Parque Nacional Cumbres de Monterrey. CONANP. https://www.fs.usda.gov/about-agency/international-programsimg/pdf/programas/cumbres_de_monterrey.pdf
36. Stocks, B. J., & Martell, D. L. (2016). Forest fire management expenditures in Canada: 1970–2013. The Forestry Chronicle, 92(03), 298–306. https://doi.org/10.5558/tfc2016-056
37. Tymstra, C., Stocks, B. J., Cai, X., & Flannigan, M. D. (2020). Wildfire management in Canada: Review, challenges and opportunities. Progress in Disaster Science, 5, 100045. https://doi.org/10.1016/j.pdisas.2019.100045
38. Western, D. (2001). Human-modified ecosystems and future evolution. Proceedings of the National Academy of Sciences, 98(10), 5458–5465. https://doi.org/10.1073/pnas.101093598
39. Wong, T.-T., & Yeh, P.-Y. (2020). Reliable Accuracy Estimates from k -Fold Cross Validation. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1586–1594. https://doi.org/10.1109/TKDE.2019.2912815
40. Xie, L., Zhang, R., Zhan, J., Li, S., Shama, A., Zhan, R., Wang, T., Lv, J., Bao, X., & Wu, R. (2022). Wildfire Risk Assessment in Liangshan Prefecture, China Based on An Integration Machine Learning Algorithm. Remote Sensing, 14(18), 4592. https://doi.org/10.3390/rs14184592

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación