IOT SEMANTIC BASED DEEP LEARNING MODELS FOR DIABETES DISEASE PREDICTION

Thesis submitted to the Bharathidasan University
in partial fulfillment of the requirements for
the award of the degree of
DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

By

S. SATHYAPRIYA, M.Sc., M.Phil.,
(BDU Ref. No. 01059/Ph.D.K10/Comp.Sci./F.T./May 2019)

Under the Guidance and Supervision of

Dr. L. AROCKIAM, M.C.A., M.Tech., M.B.A., CSM., BLIS., M.Phil., Ph.D.,
Associate Professor

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE
St. JOSEPH'S COLLEGE (Autonomous)
Special Heritage Status Awarded by UGC, Nationally Accredited at 'A++' Grade (4th Cycle) by NAAC
College with Potential for Excellence by UGC, DBT-STAR & DST-FIST Sponsored College
TIRUCHIRAPPALLI - 620 002, INDIA.

APRIL 2022

St. JOSEPH'S COLLEGE (Autonomous)

Special Heritage Status Awarded by UGC, Nationally Accredited at 'A++' Grade (4th Cycle) by NAAC
College with Potential for Excellence by UGC, DBT-STAR & DST-FIST Sponsored College

TIRUCHIRAPPALLI-620 002.

Phone: 0431-2700320, Cell: 94439 05333 Fax: 0431-2701501
E-mail: larockiam@yahoo.co.in Website: www.sjctni.edu

Dr. L. Arockiam, ***MCA., M.Tech., MBA., CSM., BLIS., M.Phil., Ph.D.,*** **Date :**
Associate Professor in Computer Science
Director of MCA

CERTIFICATE

This is to certify that the thesis entitled **"IOT SEMANTIC BASED DEEP LEARNING MODELS FOR DIABETES DISEASE PREDICTION"** submitted by **Ms. S. SATHYAPRIYA,** a research scholar in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli - 620 002, for the award of the degree of **Doctor of Philosophy in Computer Science**, is a record of original work carried out by her under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the University and in my opinion the thesis has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Date : **(Dr. L. Arockiam)**

Place: Tiruchirappalli *Research Supervisor*

S. SATHYAPRIYA
Research Scholar
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli - 620 002.

DECLARATION

I hereby declare that the work embodied in this thesis entitled **"IOT SEMANTIC BASED DEEP LEARNING MODELS FOR DIABETES DISEASE PREDICTION"**, is a research work done by me under the supervision and guidance of **Dr. L. Arockiam,** Associate Professor, Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli - 620 002, India. The thesis or any part thereof has not formed the basis for the award of any Degree, Diploma, Fellowship or any other similar titles.

Date : **(S. SATHYAPRIYA)**

Place: Tiruchirappalli *Research Scholar*

PG & RESEARCH DEPARTMENT OF COMPUTER SCIENCE

St. JOSEPH'S COLLEGE (Autonomous)

TIRUCHIRAPPALLI- 620 002

TAMILNADU, INDIA

CERTIFICATE OF PLAGIARISM CHECK

1	Name of the Research Scholar	**Ms. S. SATHYAPRIYA**
2	Course of Study	Ph.D., Computer Science
3	Title of the Thesis/Dissertation	**"IOT SEMANTIC BASED DEEP LEARNING MODELS FOR DIABETES DISEASE PREDICTION"**
4	Name of the Research Supervisor	**Dr. L. AROCKIAM**
5	Department/Institution/Research Center	PG & Research Department of Computer Science St. Joseph's College (Autonomous) Tiruchirappalli-620 002
6	Acceptable Maximum Limit	10%
7	Percentage of Similarity of Content Identified	0%
8	Software Used	OURIGINAL
9	Date of Verification	04.04.2022

Report on Plagiarism check, item with % of similarity is attached

Signature of the Research Supervisor **Signature of the Candidate**

Document Information

Analyzed document	Sathyapriya- thesis.docx (D132512799)
Submitted	2022-04-04T08:20:00.0000000
Submitted by	Dorairajan
Submitter email	manavaidorai@gmail.com
Similarity	0%
Analysis address	manavaidorai.stjct@analysis.ouriginal.com

Sources included in the report

ACKNOWLEDGEMENTS

Foremost, I owe my sincere gratitude from the depth of my heart to my respectful Research Supervisor **Dr. L. Arockiam,** Associate Professor in Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, for the continuous support to carry on my Ph.D study and research, for his motivation, enthusiasm, and immense knowledge. I am grateful for the tremendous time and energy he spent in guiding my research. I am greatly indebted to him.

Beside my research supervisor, I am greatly indebted to my Doctoral Committee members **Dr. T.N. RAVI**, Associate Professor, Department of Computer Science, Periyar E.V.R. College (Autonomous), Tiruchirappalli-620 023 and **Dr. R. JEMIMA PRIYADARSINI**, Associate Professor, Department of Computer Science, Bishop Heber College, Tiruchirappalli-620 017, for their stimulating motivation and valuable ideas.

I owe a debt of gratitude to **Rev. Dr. M. Aockiasamy Xavier S.J.**, Principal, St. Joseph's College (Autonomous), Tiruchirappalli for providing me a great research opportunity to pursue the doctoral program in this esteemed institution.

I wish to express my sincere thanks to **Prof. A. Charles**, Head and **Dr. D.P. Jeyapalan**, Former Head, Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli for providing an ambient research environment in the department. I take this opportunity to record my sincere thanks to all the faculty members in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli for their encouragement and support.

Special thanks to all my fellow research scholars and DARE (DrAros Research and Education) group for their discussions and deliberations during the course of my research.

I wish to express my special thanks to **Dr. G. VAITHEESWARAN** for providing valuable suggestion and support.

My heartfelt thanks to my Father **P. SUNDARRAJ (late)** who is the main reason to do this degree and mother **S. UMA MAHESVARI**, sisters **S. SASIREKA** and **S. SARANYA**, brother **S. GOBINATH** for their support and blessings to complete my research journey.

I also place on record, my sense of gratitude to one and all who extended their helping hand in completing this venture. Above all, I thank God, the Almighty for bestowing me with abundant grace especially when I waded through great obstacles in my life.

S. SATHYAPRIYA

ABSTRACT

Having known that the usage of Internet-based technologies and the enormous quantity of data it generates, the methods traditionally available to process those data become void. The datasets which are tough to process are known as “Big Data”. Such data would majorly be unstructured data; however, structured kinds of data will be limited and domain-specific, which are often found operated under IoT environments. These structured data only have an enormous quantity of useful information, that might be insightful while many decision-making processes. These unstructured data would become insightful only if those data could be processed effectively. One such effective method of acquiring insightful information is the deployment of semantics and applying ontology data into the semantics.

This difficulty of handling data would be even more difficult in the case of Health care applications, especially in effectively handling the data in case of critically dangerous diseases like Diabetes. Furthermore, integration of IoT systems and health care applications also becomes tedious. Thus, this research work is proposed by concentrating on **IoT semantic ontology**, **deep learning methods** and **domain-based knowledge**. Models are proposed and compared in each phase of the research works by evaluating the critically important performance metrics with the deployment of diverse tools.

The proposed methodology of Rule Embedded Semantic Ontology-based DNN Classifier (RESOC_DNN) was to hold together the IoT and domain-specific data. This method generates various rules with the establishment of a rule engine for

the sake of validation of ontology, which is further aided in developing the classification of the diabetic disease. First, diverse series of data was collected from various IoT enabled devices. Then, the ontology is built and the rules for the prediction is framed, followed by the building of the DNN-based model. Finally, the classification of data into two categories such as “Diabetic” or “Non-Diabetic” takes place with considerable accuracy.

The proposed methodology of Improved-DNN with binary Multi-objective Optimization (IMO_DNN) for predicting the risk of Diabetes Patients was done to realize further risk levels of the Diabetic disease along with the prediction of disease to prevent the patients from death or ending up in a more critical condition. The prepared ontology data would be given as the input to predict Diabetes mellitus with the integration of the loss function, Softmax function, and Binary Multiple-Objective Optimization for categorizing the risk levels into three, namely: “Low Risk”, Medium Risk” and “High Risk”. The objective of the work was to achieve better performance characteristics in terms of accuracy, precision, F1 Score, and Recall, which was accomplished and compared with other machine learning methodologies like RNN, CNN, and ANN.

The proposed methodology of an Enhanced IMO_DNN (EIMOXB_DNN) with Boosting was done for improving the accuracy of risk prediction categorization made in the previous phases. The capabilities of categorizing the data into low or medium or high risk is improved by using the concept of IMO_DNN with methods of XGBoost, feature selection, Classification, and Adam Optimization. This proposed model ensured that the diabetes disease risk levels were properly predicted with the optimized features. Like the previous work, the performance of the Enhanced IMO_DNN model was evaluated in terms of accuracy, precision, F1Score, and Recall by comparing against the models like IMO_DNN, RNN, CNN and ANN.

The more improvised model was Enhanced IMO_DNN, which was able to predict diabetes disease along with the risk levels with better accuracy and precision.

PAPERS PUBLISHED

1. **Sathyapriya, S**., and L. Arockiam, “Rule Embedded Semantic Ontology Based Classifier For IoT Healthcare”, *Annals of the Romanian Society for Cell Biology*, ISSN: 1583-6258, Vol. 25, Issue 4, 2021, pp. 10224-10231. **(Scopus Indexed)**
2. **Sathyapriya, S**., and L. Arockiam, “A Review On semantic Technology in IoT Healthcare”, *Annals of the Romanian Society for Cell Biology*, ISSN: 1583-6258, Vol. 25, Issue 4, 2021, pp. 9810-9819. **(Scopus Indexed)**
3. Dr. L. Arockiam, **S. Sathyapriya**, V. A. Jane, A. Dalvin Vinoth Kumar, “Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India”, *International Journal of Recent Technology and Engineering*, ISSN: 2277-3878, Vol. 8, Issue 2, July 2019, pp. 6440-6444. **(Scopus Indexed**).
4. Dr. L. Arockiam, **S. Sathyapriya,** V. A. Jane, A. Dalvin Vinoth Kumar, “Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning”, *IJRTE Blue Eyes Intelligence Engineering & Sciences Publication*, ISSN: 2277-3878, Vol. 8, No. 2, July 2019, pp. 6400-6403. **(Scopus Indexed)**.
5. P. Kamakshi Priyaa, **S. Sathyapriya,** L. Arockiam, “Nutrition Monitoring and Calorie Estimation using Internet of Things (IoT)”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, Vol. 8, Issue 11, September 2019, pp. 2669-2672. **(Scopus Indexed)**.
6. P. Kamakshi Priyaa, **S. Sathyapriya,** L. Arockiam, “An Enhanced Decision Tree Ensemble Technique for Obesity Prediction”, *International Journal of Scientific & Technology Research*, ISSN: 2277-8616, Volume 9, Issue 2, February 2020, pp. 1770-1775. **(Scopus Indexed)**.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	- Artificial Intelligence
ANNs	- Artificial Neural Networks
API	- Application programming interface
APSO	- Adaptive Particle Swam Optimization
BP	- Back Propagation
CNN	- Convoluted Neural Network
COPD	- Chronic Obstructive Pulmonary Disease
CUIs	- Concept Unique Identifiers
DL	- Deep Learning
DM	- Data Mining
DMLC	- Distributed Machine Learning Community
EAA	- Enhanced Apriori Algorithm
EHR	- Electronic Health Record
EHRs	- Electronic Health Records
FC	- Fully Integrated
FHIR	- Fast Healthcare Interoperability Resources
GNU	- GNU's Not UNIX
GUI	- Graphical User Interface
GWO	- Grey Wolf Optimization
HAR	- Human Activity Recognition
HbA1c	- Hemoglobin A1c
HTTP	- Hypertext Transfer Protocol

IDE	- Integrated Development Environment
IoHT	- Internet of Healthcare Things
IoT	- Internet of Things
JSON	- JavaScript Object Notation
KNN	- k-Nearest Neighbors
LR	- Logistic regression
LSTM	- Long short-term memory
MATLAB	- MATrix LABoratory
ML	- Machine Learning
MLP	- Multilayer Perceptron
MPI	- Master Patient Indexes
MQTT	- Message Querying Telemetry Transport
NER	- Named Entity Recognition
NN	- Neural Network
PHRs	- Personal Health Records
RDF	- Resource Description Framework
ReLU	- Rectified linear unit
RNN	- Recurrent Neural Network
SAREF	- Smart Appliances REFerence
SMO	- Sequential Minimal Optimization
SMO	- Sequential minimal optimization
SSN	- Semantic Sensor Network
SVC	- Support Vector Classification
SVM	- Support Vector Machine

SWRL	- Semantic Web Rule Language
UMLS	- Unified Medical Language System
URL	- Uniform Resource Locator
WSN	- Wireless Sensor Network
XGBoost	- eXtreme Gradient Boosting
XML	- Extensible Markup Language

Chapter – 1

Introduction

CHAPTER - 1

INTRODUCTION

1 Introduction

The introductory chapter of the thesis discusses the basics and applications of Internet of Things (IoT) in the sector of Heath care for the sake of diagnosing various diseases or ailment towards the betterment of the Human health. Then, the functions and evolution of IoT along with its layers present in it is discussed.

Subsequently, the basic concepts of Data Analytics, after which the relationship among IoT and Big Data is described, techniques to process the IoT data and applications are briefed.

Then, the basic concepts of IoT semantics are discussed, after which the need, significance of semantics in IoT Application, and a few of the earlier proposed Semantic-oriented process are briefed.

Then, the basic concepts of Deep Learning are presented with the help of its application in IoT arena. Various classification methodologies proposed based on Deep Neural Networks are also briefed.

Next, the background and motivation pertaining to the present work is presented to identify the Research Gaps / Problem Statements. After identifying the Problem Statement, the objectives are projected to address the various limitations in the health care. At the last, the thesis organization is written, wherein all the five chapters are outlined and summarized.

This work is majorly proposed and implemented with background concepts of IoT-oriented methods to diagnose the Diabetes diseases.

1.1 Introduction to Internet of Things (IoT)

Internet of Things (IoT) refers to a number of "things" that are interconnected with the utilization of Internet for the sake of sharing information among all other things connected [Pat, 16]. Different types of IoT based applications were found deployed in diversified fields, which had been successful in producing favorable benefits to its users. These devices generate enormous amounts of data every day, and this data can only become valuable if it could be analyzed to produce insights [Da, 14].

In all the sectors, various devices need to be involved for applications like measuring or monitoring or identifying any type of condition. Thus, the IoT becomes the most sought solution for many such sectors as found in the below fig. 1.1 in order to facilitate the involvement of huge number of devices even in the most complex type of systems.

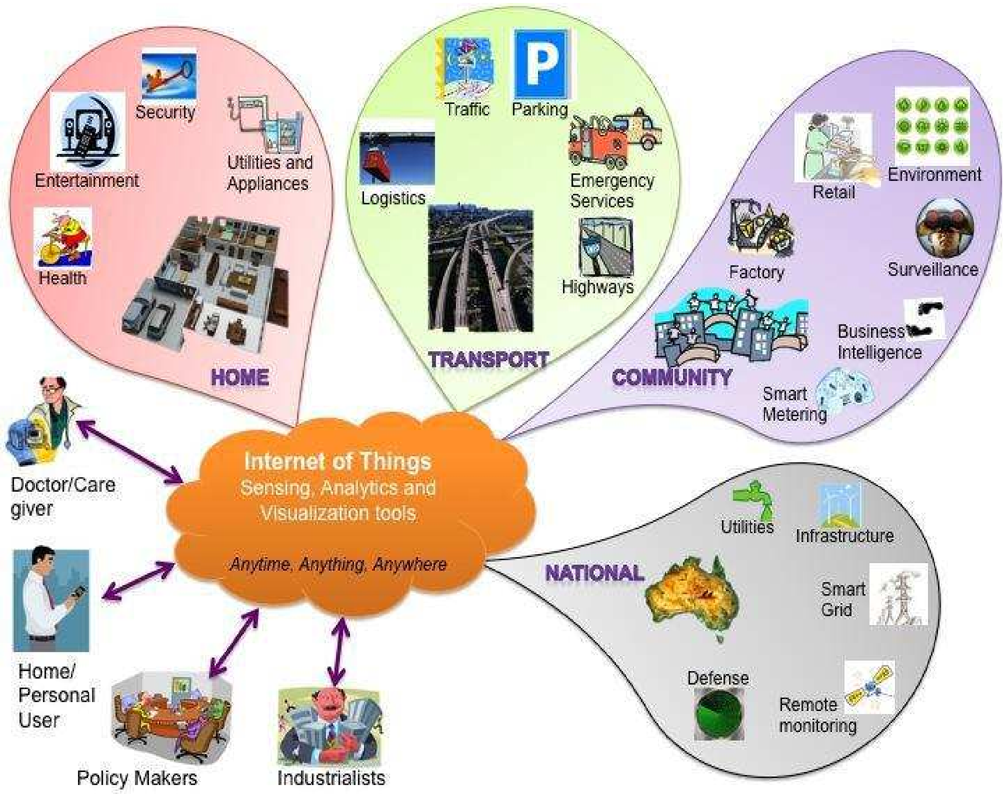


Fig. 1.1 Sector Wise-application of IoT [Gub, 13]

1.1.1 Functions of IoT

Various functions of IoT are as follows:

- A device can be connected with each other irrespective of the configurations that may vary from one device to another device.
- There will be no limitation in the number of devices connected to the Internet.
- Any user irrespective of the count can monitor and maintain various data gathered by all the interconnected devices.
- The devices in the IoT environment can be interconnected with each other through any paths or networks.
- The location of the device and user is not the constraint. Thus, the user can operate or monitor or maintain the interconnected devices from anywhere.
- As [Gub, 13] says, IoT is found deployed in various sectors incuduing health care, education, and agriculture.

The following fig 1.2 is an depicts the various funtions of a classic IoT environment.

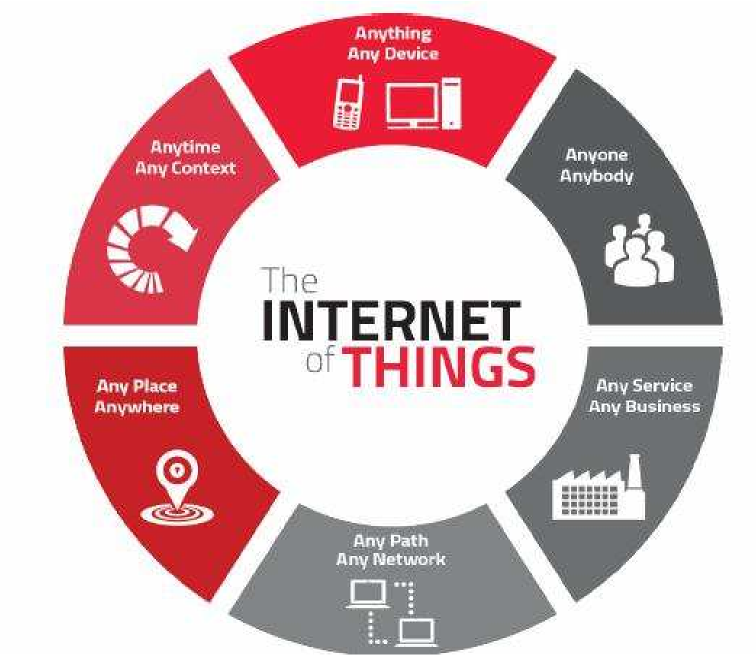


Fig. 1.2 Functions of a Typical IoT Environment [Pat, 16]

Thus, IoT application may possibly be actuated or operated from any place with less or nil constraints [Pat,16].

1.1.2 Layers of IoT

Any particular IoT environment may be powered with the help of different layers in it depending upon its range of application. Likewise, one such IoT architecture as per [Pat,16] is taken and summarized which is depicted in Fig 1.3. According to them, there are four layers, which are as follows:

- **Application Layer:** In this Application layer of IoT, many functionalities like Monitoring, Predictive/ analytics, and context-aware computer service, etc. were carried out [Pat, 16].

- **Service support and Application support Layer:** This is a supporting layer that follows the application layer that aids to function smoothly without any issues. The layer is responsible for functions like: Authentication, Identity Management, Business Process Modelling, Encryption, Data Quality Management, and Virtual Entity, etc. [Pat, 16].

- **Network/ Communication Layer:** This layer has the capability to establish the connection between the application layer and sensor layer by using communication sharing methods. For instance, the communication sharing methods like LTE, Wi-Fi, and Ethernet, etc. to enable give rise to better Network as well as Transport capabilities [Pat, 16]. However, the application layer for the sector like Health care and education will be different.

- **Smart device / Sensor Layer: Smart Device/ Sensor Layer:** This is the last layer which normally consists of smart devices that are either wireless like Wi-Fi and Bluetooth or either wired like Ethernet and other related components. Moreover, it also possesses many devices such as analog devices, digital devices, gyroscope, etc. which can be utilized in health care sector.

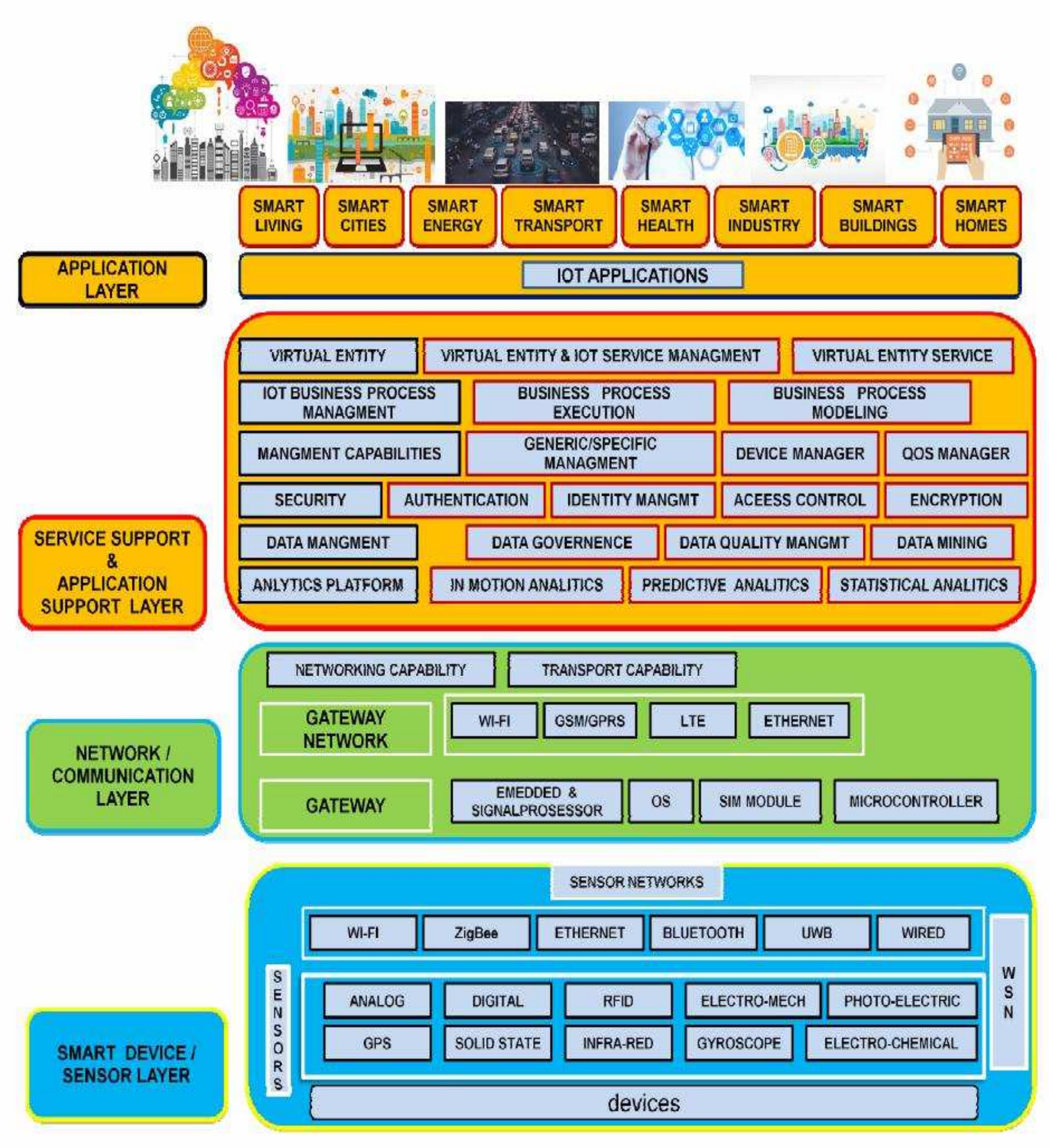


Fig. 1.3 Applications of IoT with respects to its layers [Pat,16]

1.1.3 Evolution of IoT

The IoT was extensively used in the wireless communication arena, where it evolved a lot right from Routing to smart cities [Can, 18]. But, nowadays, the IoT had been started using in the health care, wherein many diseases/ ailments are diagnosed in an early stage to avoid unforeseen medical difficulties or causalities. However , the skills required keep-up with this evolution of IoT in health care was found to be difficult and it needs much attention [Mar, 18].

IoT application has not only evolved in terms of architecture, but it had also evolved in terms of its security [Rom, 18]. Thus, when applying it to a new sector like health care, the data security of the patient and diagnosis data needs to be taken care of.

1.1.4 IoT in Healthcare

Among various applications of IoT, Healthcare domain has driven more attention. IoT technologies provide efficient service to the healthcare domain like monitoring patient, diagnosing, giving treatment, taking a quick decision, minimizing the cost, and avoiding the critical issue [Kar, 15]. In general, IoT offers wearable devices for patient monitoring and making decision about the current status of the patient. Nowadays, IoT medical sensors are found to use widely, where it generates big volume of health-related data [Mou,16]. This data would be analyzed carefully because decisions are very important in the healthcare domain.

But, one of the major issues in this was data formats, because data were collected from various sensors and from various sources. So, this will lead to heterogeneity, interoperability and scalability issues. To avoid these issues, semantic web technologies will be the go-to option to provide a unique data model to gather data from various sensors [Mav,19]. Normally, it was used to convert the data to meaningful form and explore the structures and relationships between the data.

1.2 Data Analytics

1.2.1 Basic Concepts and Definition

Normally, data analytics process is break down into numerous stages like preparation of data, preprocessing of data, data analysis and post processing of data. These different phases involved in the data analytic process are represented in the fig. 1.4. Let's look onto this topic in brief manner.

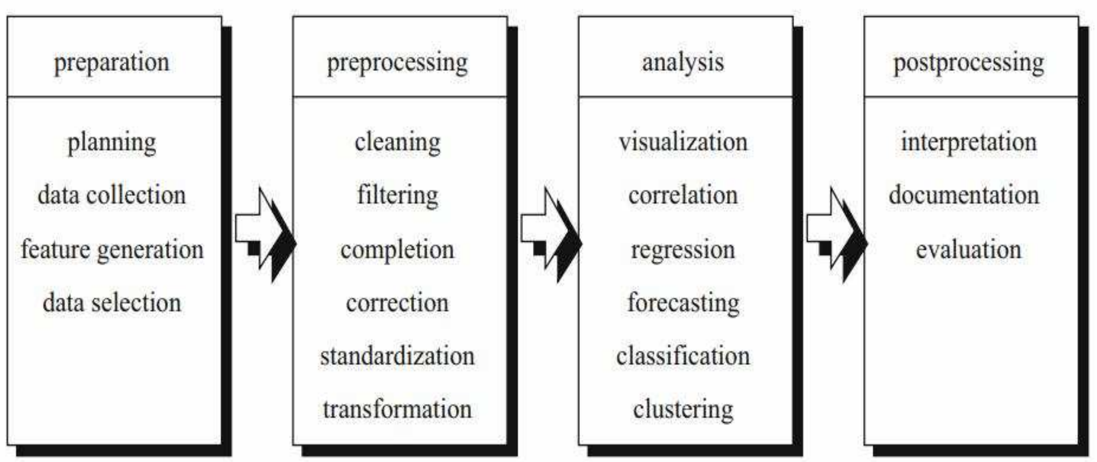


Fig. 1.4 Phases of Data Analysis [Run, 20]

a. Preparation of data

It is the procedure of purifying and reforming source data former to examining. This is one of the significant phases before processing. Further, it involves redesigning data, making corrections in the data, and incorporating data packages to enrich the data. Data production is often a long-term endeavor for data experts or industry users, but placing data in an environment is important as a precondition for converting data into knowledge and removing dependencies caused by declining data quality.

b. Preprocessing of data

The task of converting source data into a reasonable format is known as data preprocessing. This is a significant phase in data analytics because source data contains many defects such as missing values, noise values, etc., Data quality should be verified before using Data Mining (DM) or Machine Learning (ML) techniques.

c. Data analysis

The process of analyzing large amount of data in order to take a certain decision is defined as data analytics [Run, 20]. It is done through the help of computer applications. It makes use of various other interdisciplinary concepts like artificial intelligence, machine learning, statistics, pattern recognition, etc.

d. Post processing of data

This method generally contains several alignment procedures, rule filtration or even information combination. All of these events offer a type of filter for inaccurate and noisy information obtained by an analytic algorithm. So, some pre-processing procedures and latter processing should load the complete sequence of data Processing.

1.2.2 Relationship between Big Data and IoT

Internet of Things (IoT) is capable of generating Big Data for numerous reasons [OL, 13]. A considerable volume of data is attributed to the IoT. When various sensors that are capable of capturing data continuously, interact with things like Radio Frequency Identification (RFID) tags huge volume of data are generated. The types of data related to the IoT are based on certain factors like types of sensors and various sources of data.

1.2.3 Techniques to Process the IoT Data

IoT data are processed by using different types of techniques. Some of the techniques that are used for processing IoT data are mentioned below [Kri,20]:

- Data Denoising
- Data Classification and Prediction
- Data Outlier Detection
- Data Sorting
- Missing Data Imputation
- Data Calculations
- Data Aggregation

1.2.4 IoT Data Analytics Applications

Application of IoT Data Analytics include various government and private organizations or institutions where the applicant's information are verified at the time of recruiting [Sho, 18].The details of the citizens are stored in g-cloud, through which verification can be done by easily retrieving data from the cloud. This process is carried out through deep neural networks which demand the proper analysis of big data.

1.3 IoT Semantics

1.3.1 Basic Concepts

The word 'Semantic' refers to the meaning of data which describes about a single data in a detailed manner which results in a better and easier communication through offering interoperability among devices [Bur, 19]. Resource Description Framework (RDF) is a major technology which is employed to afford a separate platform for semantic model that enables interoperability among different IoT devices [Ban, 18].

1.3.2 Need of Semantics in IoT Application

The need of semantics is a most significant one in IoT as it plays a key role in the knowledge system [Wan, 13], as it aids in data enrichment, handling heterogeneous materials, interoperability, and gaining necessary knowledge from the IoT information. For better understanding of a data, it represents single measurement data with various metadata like the description of sensors, context and configuration of the sensors [Ber, 17]. Semantic modeling generates a clear understanding of a data in a more structured manner by combining context related information and domain knowledge with raw information. Semantic web technology initiates the devices to know more about the processing of information through machine readable and structured descriptions of data. Descriptions like type, location, relationship are existing among the context and domain, data provider and various data attributes [Pal, 16].

1.3.3 Ontology

The term ontology is often used in the field of philosophy which is related to the existence of subjects and matters. It indicates the specification of conceptualization. Ontology is described as the study of concepts like existence, reality and of being or becoming one. It illustrates about the relationship and concepts of subjects which exist for a matter or group of matters. It is also defined as group of subject definitions which are consistent and general [Gru, 18]. Ontology usage in IoT semantics is completely different than to its usage in the field of philosophy.

Ontology plays a major role in semantics as it gives details like data characteristics and explanations [Seo, 19]. Ontology modeling concept is currently in an improvement stage that is continuously improving by every day and is undergoing sustainable development gradually. Usually rich and complicated data are involved in Ontology [Che, 20].

1.3.4 Semantic Approaches

Semantic approaches are classified to three different types. They are:

- Linguistic Approaches
- Ontology Based Approaches
- Mixed Mode Approaches

The above types of semantic approaches can be further subdivided into different types based on their key features. The classification of semantic approaches is depicted in the below fig 1.5 which are explained briefly in this section. [Raz, 19]

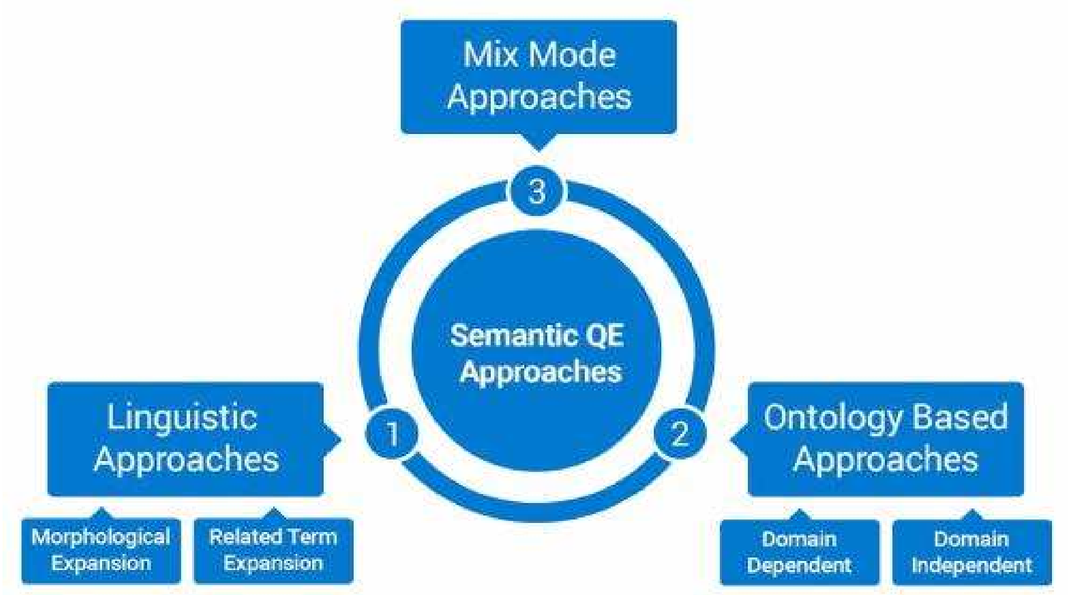


Fig. 1.5 Classification of Semantic Approaches

a. Linguistics Approaches

Latest semantic approaches concentrate more on term dependency of data in natural language like morphology or related terms for the expansion of initial search query. Since, these approaches exploit about the natural language properties for the generation of key expansion terms these are known with the name of linguistic approaches. Linguistic approaches are further classified into:

- Morphological Expansion
- Related Term Expansion

b. Ontology Based Approaches

In this type of approach, ontology concepts are added to the initial queries before the submission of queries. This concept indicates the subjects present in a domain and are semantically linked for forming a knowledge-based structure known as ontology. Hence, this approach is known with the name of ontology-based approach. Ontology approaches are further classified into:

- Domain dependent
- Domain independent

c. Mixed Mode Approaches

Mixed mode approach has both the characteristics of linguistic and ontology-based approaches. Hence this approach is known with the name of mixed mode approach. Researchers have performed several investigations combining by both the lexicon (linguistic) and ontology (semantic) techniques.

1.4 Deep learning

Deep Learning is a sub division of the Machine Learning technique which has witnessed a rapid growth over the last 6 years or so. Deep learning is purely directed by the availability of large datasets and the enhancement in the computational power. The deep learning concept has observed improvements and advancements in the machinability for understanding and manipulating the available data like language, image and voice commands. Nowadays the healthcare and medical field are benefited more from the deep learning techniques due to the huge volume of data that is being generated (e.g., 1018 bytes or 150 Exabyte's in United States of America alone and results in the 48% growth every year). It also increases the propagation of digital record systems and proliferation of medical devices [Est, 19]. The fig 1.6 is for representing the concept of Deep Learning.

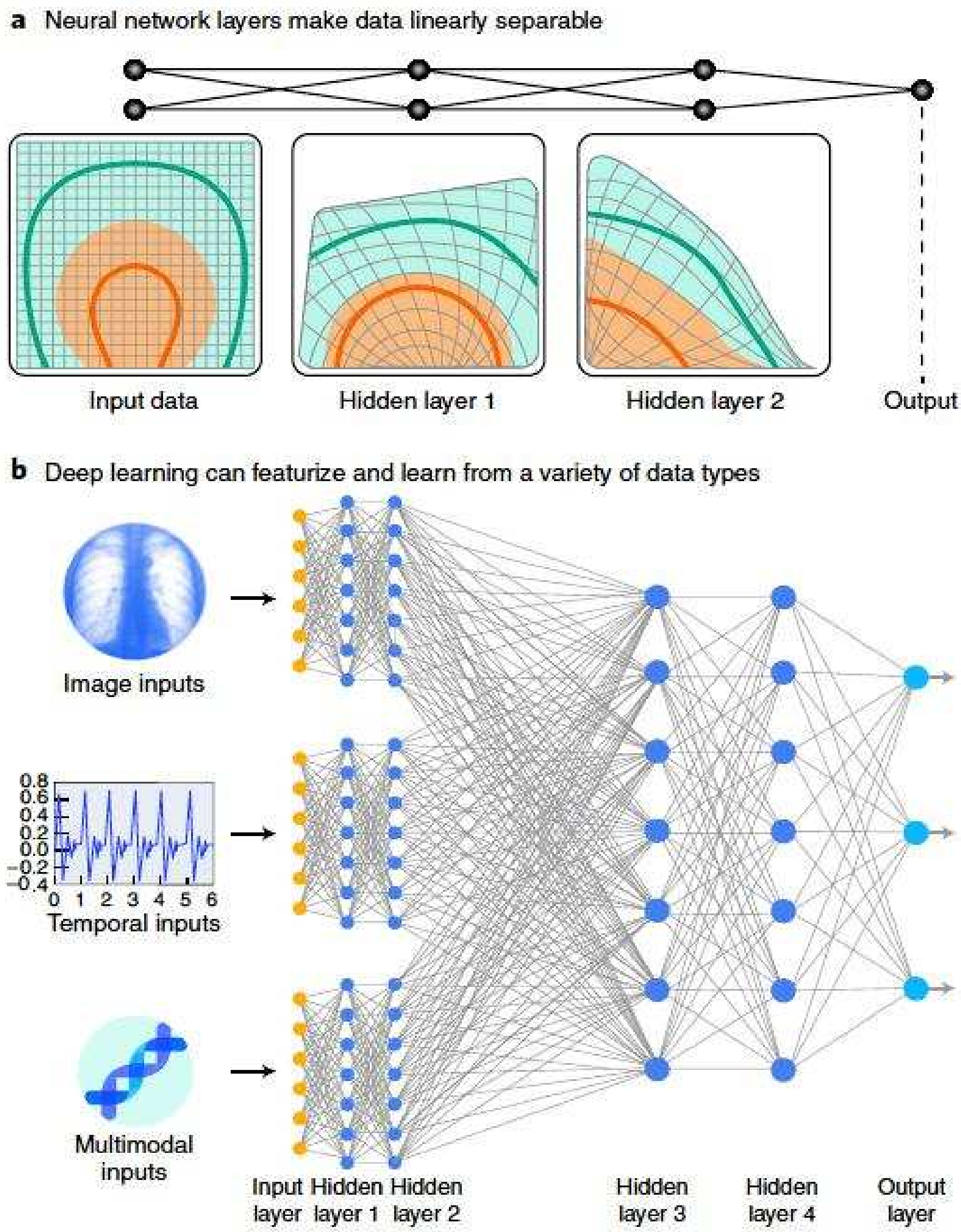


Fig. 1.6 Representation of Deep Learning

1.4.1 Classification of Deep Learning Techniques

Deep learning [Ati, 20] can be classified into different categories like:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

Moreover, a typical representation of each category is depicted in fig 1.7.

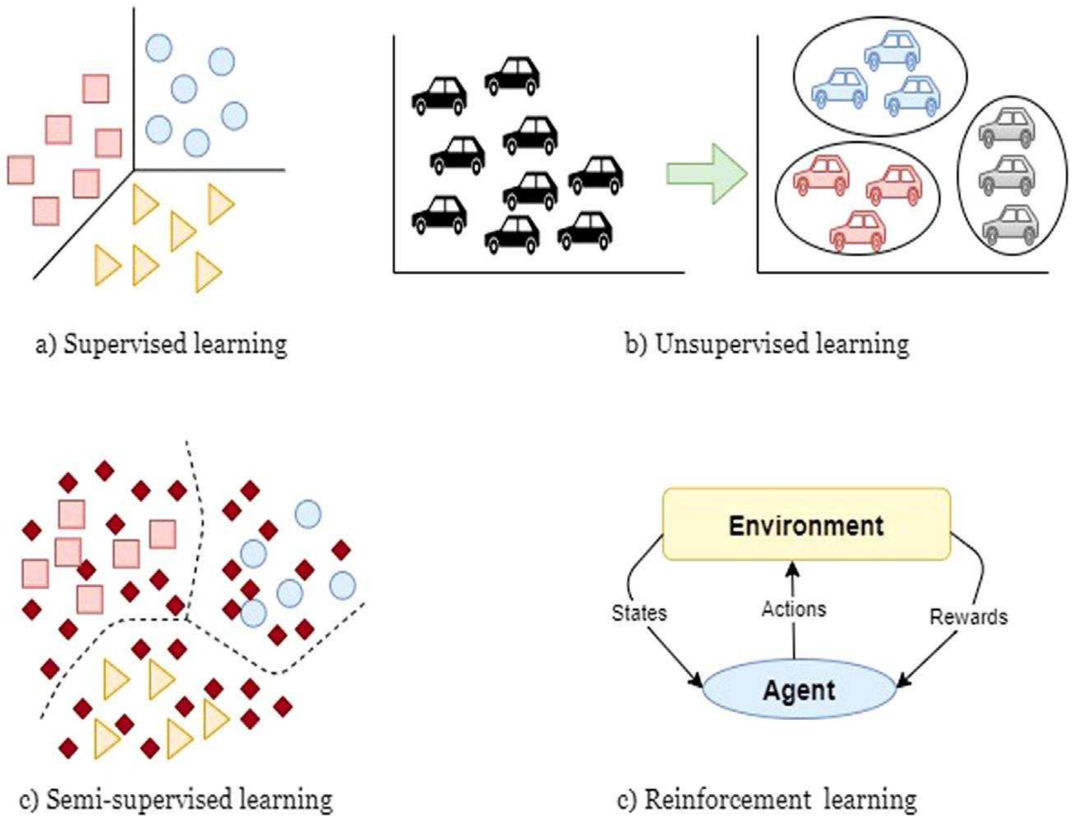


Fig. 1.7 Classification of Deep Learning Techniques

1.5 Significance of Predicting Diabetes Disease

The significance of knowing about the rate of diabetes in our human body could help us avoid any critical medical conditions like blindness, kidney failure, and any related abnormalities. Determination of HbA1c values plays a crucial role in aiding the diabetes diseased person. Some of the major significance of predicting the level of diabetes diseases [She, 16] are as follows:

- Early predicted higher values of HbA1c can help the diseased person to prevent cardiovascular risks and strokes [Mar, 14].
- The early predicted higher values of HbA1c also avoids unnecessary death of the diseased person, thereby reducing the mortality rates [Sel, 10].
- The early predicted lower values of HbA1c can help the person to know the risks towards the condition of hypoglycemia.

1.6 Research Motivation

As many of the IoT methods adopted in the health care sector are very recent and rare, the issues like heterogeneity, interoperability, and scalability issues are known to occur. Even, many researchers consider this as an uncertainty in their works while deducing the results. However, elimination of all other issues like heterogeneity, interoperability, and scalability issues, are not easy and thus more reliable type of advanced IoT-oriented prediction models will be desired.

1.7 Problem Statement

- Internet of Things is a blooming technology in the modern world including the health care. Lots of IoT devices connected through Internet communicate with each other. Data from multiple sources poses a key challenge to create actionable information from raw data for developers and researchers alike. These enormous amounts of data can only become valuable if it is analyzed.
- Data analysis is the process of analyzing large and small data sets with different data properties to extract meaningful results and active insights. These results are usually in the form of patterns that help to make an effective and quick decision in all domains.

- Early detection of the effects of diseases can reduce the risk in healthcare domain. In the past era, many researchers have been doing their research in predicting the diseases in healthcare domain. But still an effective model is still needed to predict accurately and take timely decisions by using improvised classification methodology.

1.8 Scope of the research

IoT analytics is a new and challenging field. It takes in huge volumes of heterogeneous data from IoT devices, with the objective of processing, storing, and extracting business value from them. Semantic based model improves the system to handle heterogeneous data from IoT devices in the current research field. Moreover, IoT semantic data are meaningless without extracting knowledge out of it, so there is a need for analytics. Many analytical models have been developed but an accurate model is needed to critically analyze the semantically enriched data. So, there is a need to provide model to manage heterogeneous data and analyze accurately for gathering insights from it. Scope of the research work is narrowed down to propose models for handling heterogeneity and then accurately predict and classify valuable insights out of it.

1.9 Objectives of the work

The present work desires to overcome all the possible difficulties in diagnosing the diabetes diseases by deducing the following objectives:

- To propose ontology based deep learning model for IoT data to avoid the heterogeneity issues and improve accuracy of classification
- To build a new deep learning model using optimization functions and clustering to enhance the decision making
- To propose an enhanced model to increase the accuracy of prediction
- To prove the better performance of the current work with the existing works.

1.10 Thesis Organization

This present work documented in the thesis is organized and presented in different chapters. Summary of each chapter are discussed below.

Chapter 1 gives a brief introduction on the meaning of Internet of Things (IoT) along with layers, functions, and evolution. It also briefs about the data analytics, semantic ontology and briefs on the basics of deep learning methodology and classification with reference to IoT Analytics. Then, it discusses about the significance of diagnosing the diabetes diseases, motivation behind the research in consideration, and problem statement of the presented work. Finally, the objectives of the work are presented by mapping the problem statements identified.

Chapter 2 reviews the basic terminologies of the existing such as: IoT; Data Analytics; IoT Semantics; Deep learning; and Detection of Diabetes disease. Furthermore, Deep learning methodology for IoT and Diabetes disease prediction with the help of IoT will be extensively detailed with the help of earlier presented literatures.

Chapter 3 discusses the research objectives and method of research work.

Chapter 4 discusses two phases like semantic development phase and prediction phase in the Rule Embedded Semantic Ontology Based DNN Classifier Model regarding the conversion of data into meaningful and understandable format for quick analysis. The performance measures are compared at the last.

Chapter 5 discusses about the IMO_DNN model for predicting the risk of diabetes patients by trained deep layers. The performance measures are compared at the last.

Chapter 6 discusses about the Enhanced IMO_DNN with Extreme Gradient Boosting adopted for improving the speed and performance of the prediction model.

Chapter 7 illustrates the presented research work and suggests for future research.

Chapter – 2

Literature Review

CHAPTER - 2

LITERATURE REVIEW

This chapter discusses the various works found in the literature pertaining to the IoT semantic, Data Analytics, Deep learning based prediction models and Diabetes Disease Prediction. The chapter is categorized into seven sub-sections, starting with Introduction, which briefs about the various sub-sections in the chapter.

2.1 Introduction

First of all, to start, the basic concepts and definitions pertaining to the various major terminologies such as the present work of IoT, Data Analytics, IoT Semantics, Deep Learning, and Diabetes Disease Prediction, etc. are discussed one by one.

Then, the IoT Semantics Techniques will be surveyed and briefed with reference to Ontology for IoT and Healthcare along with the Analytical Survey. Then, the Deep learning methods for IoT environment will be surveyed along with its extensive applications found in the sector of Health care.

After giving a basic idea of IoT Semantics Techniques and its applications, the importance of predicting Diabetes Disease in Human beings is summarized, which is capable to prevent and avoid any unforeseen critical medical conditions / severity.

Many of the advanced technologies were being deployed by earlier researchers to successfully predict the Diabetes. However, there were limitation in it [Sun, 20]. Finally, various issues and challenges faced while implementing such IoT Semantic based Deep Learning Models for Diabetes Disease Prediction is discussed and chapter summary is presented after it.

2.2 Basic Concepts and Definitions

No prediction methods can be developed without involving some fundamental and basic concepts. Likewise, some of the basic concepts and definitions pertaining to the IoT Semantic based Deep Learning Models for Diabetes Disease Prediction are presented below.

2.2.1 IoT

IoT is nothing but the collection of physical objects like software, sensors, and other technologically advanced gadgets, that are interconnect with each other on Web in order to facilitate the process of data exchange taking place between each other in any particular system irrespective of its application [Sha, 20].

Ray et al., [Ray, 16] reviewed IoT technologies for several domain based services like device and system management, heterogeneity and data management, tools and visualization. Additionally, various challenges were discussed that were faced by many researchers.

Likewise, Samie et al., [Sam, 16] presented a survey on IoT technologies and described the characteristics related to IoT. Various important technologies which was used mainly for developing IoT based systems were illustrated for researchers who are interested to work with IoT. Also, Current trends and research issues were explained.

Furthermore, Wamba et al., [Wam, 13] discussed a classic IoT setting that includes sensors, interfaces, sophisticated techniques, and cloud technology. Sensors were played a major part in data collection from different devices. In addition, advanced technologies like RFID, WSN network and communication were offered. Also, recent methods were utilized to data analyse and process.

In [Asg, 19] [Sul, 14] Multiple client / server requirements could be exchanged in the cloud and permitted client to access different kinds of services concurrently. Fog computing was developed because of the various issues of cloud computing like delays, dependability, and source limitation, etc., Also, for real time application processing and effective decision making at anywhere [Far, 18], [Sof, 21]. In spite of all the enhancements in IoT, the new model in immaturity level and covered numerous research areas on diverse problems namely benchmarking, device diversity, scaling, security, privacy, etc., During, the interoperability of elegant devices to converse and connect diverse devices were easy to implement and inexpensive [Jab, 17]. Additionally, short expenditure capability between elegant devices is an important characteristic for IoT, which enables customers to go on working with dissimilar dealers in the future. Lastly, IoT could diminish the rate of constructions, resolve the problem of managerial communications and assist in maintaining heterogeneous communications [Ull, 17], [Naj, 21]. Currently, the IoT prototype involves a diversity of domains.

The existence of IoT was able to be possible till now with the help of following technologies [Nor, 19]:

- Cloud Computing capable Platforms
- Cost effective low powered Sensor
- Connectivity with wide pool of Network Protocols
- Conversational AI-Artificial Intelligence
- Machine Learning and Analytics (Now-a-days, evolved as Deep learning)

2.2.2 Data Analytics

Data Analytics is the method of analyzing raw datasets, resulted out of various data gathered in appropriate methods, to draw insights such as trends, conclusions, and identifying the areas of improvement [Pit, 21].

Likewise, data analytics in health care was found to use historical as well as present data for the sake of gaining insights, micro and macro, and support decision-making in both the patient and business levels [Zak, 15]. Some of the tools used for health care are:

- EHRs-Electronic Health Records
- E-prescribing-Electronic Prescription Services
- MPI-Master Patient Indexes
- PHRs-Personal Health Records
- Patient Portals

Rehman et al., [Reh, 19] presented a concept of big data analytics in Internet of Things. The fundamentals of big data analytics with its features were discussed. The significance of data analytics in IoT environment and several data analytics based on IoT were described.

Mittal & Sangwan [Mit, 19], provided a comprehensive review of data analytics tools and techniques. Mainly, described the machine learning technique with its categorization such as supervised, unsupervised and reinforcement learning.

Ghani et al. [Gha, 19] surveyed the various big data analytics techniques. Particular, some technique were described clearly that were deep learning, artificial neural networks, swarm intelligence, fuzzy systems and evolutionary computation. These techniques were evaluated by the quality metrics.

Saleem [Sal, 21] et al., presented the concept of IoT and big data analytics with its applications. The author explained the role of data analytics in IoT domains, and created taxonomy for IoT based data analysis. Various applications like healthcare, agriculture, education etc., were explained with analytical techniques.

2.2.3 IoT Semantics

IoT Semantics is the process of interpreting a detailed meaning of data handled in an IoT-enabled system [Ahl, 20]. It is generally used to resolve the issues of ambiguity and heterogeneity in the vastly gathered data in any application including IoT. Below figure shows the Typical Semantic IoT layer.

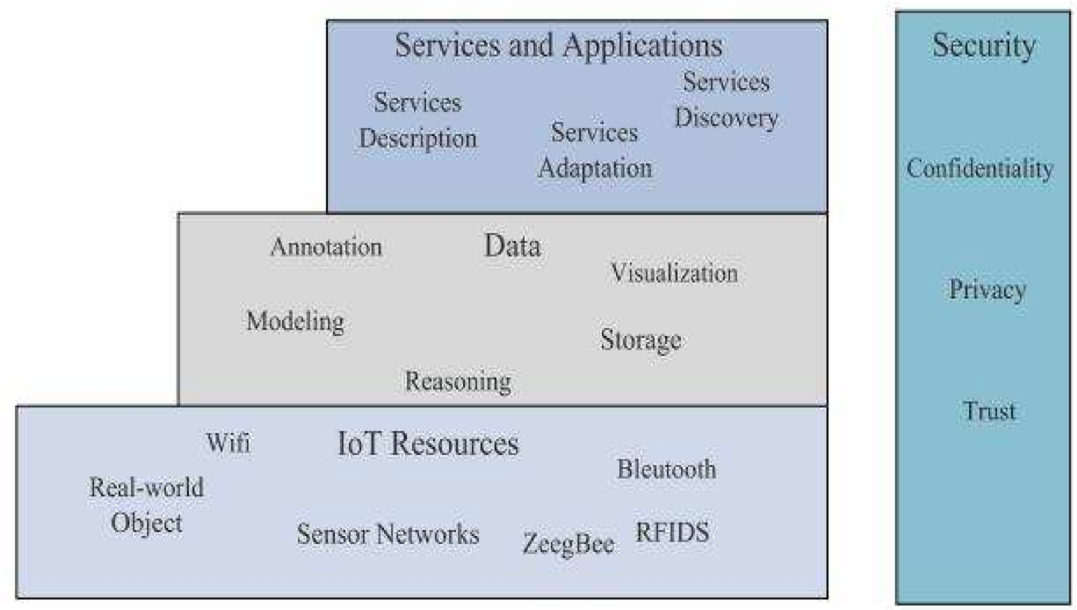


Fig 2.1 Typical Semantic layer in IoT environment [Bar, 12]

A typical semantic layer in IoT environment was depicted in fig 2.1. Let's look on to an instance where W3C SSN ontology was responsible for sensor device description and resource annotation, only when the nodes are properly linked to a gateway. Some of the recent studies which offer efficient representation of semantic data for IoT environment include the designing of various lightweight semantic description models and the designing of efficient representation models like Binary RDF representation.

Honti et al., [Hon, 19], explained the semantic sensor concepts for IoT based architectures. Several related concepts were discussed and drawn information. Several ontologies that were most commonly utilized by various researchers were summarized. However, the author said there was a need for attaining better interoperability and flexibility.

Mervin et al., [Mer, 21], surveyed the semantic web technologies for Internet of things Environment. Also, the author discussed various research works which were closely relevant to the topic. Several challenges and future directions in the area of IoT semantic were described.

2.2.4 Deep learning

Deep Learning is the evolved from machine learning methodology that is typically a neural network comprising of two or more layers in it for simulating the characteristics of human brains when dealing a vast amount of data [IBM, 21]. The neural layers empower to match to the capabilities of human brains, thus allowing to learn. When a single layer could deal with vast data for various prediction applications, many hidden layers were found comprised in it to optimize and increase its accuracy.

Mohammadi et al., [Moh, 18] surveyed deep learning techniques for IoT based applications. The author explored the recent types of DL techniques that were used in several IoT based applications. Also, the challenges of deep learning and its research gaps were explained in IoT environment.

Wanke et al. [Wan, 18] established deep neural networks concepts that could be improved by applying the information gained through the visualization of output data obtained in each layer. Several studies had been focused on visualization scheme improving in the neural networks by neural network techniques.

2.2.5 Diabetes Disease Prediction

Diabetes is a chronic medical severity in the pancreatic area in the human body, which doesn't produce insulin effectively or its utilization is not effective [WHO, 21].

According to [WHO, 21], the World Health Assembly came forward to strengthen the control and prevention of diabetes diseases having known that the Diabetes was the ninth highly death causing disease. For control and prevention of diabetes, the prediction of the disease could aid one to avoid many risk factors including unnecessary death. Thus, the Diabetes Disease Prediction becomes indispensable.

The prediction of the said disease could be possible right from the mere determination of HbA1c to the recently advanced classification methodology.

AIZubi et al., [AIZ, 20] proposed a model for diagnosing diabetes disease using the technologies of big data and machine learning. Initially, data was gathered and combined through MapReduce model. Then, the process of normalization was utilized to reduce the noise in the collected data. Furthermore, the ant bee colony algorithm was applied to choose features and made these features efficient by utilizing SVM with a multi-layer neural network. To classify the learned features, a corresponding neural network was utilized.

Chen et al., [Che, 19] created ontology for diabetes disease prediction. It was developed by utilizing medical data. Then the illness was identified and medicines were recommended. But, the proposed ontology was not suitable for IoT based applications.

2.3 IoT Semantics Techniques

As the number of interconnected devices deployed in any content of IoT environment is increasing day-by-day, the possibility of interpretation of the data handled being is higher [Ahl, 20]. As a result, the source of information understood by the recipient device might be different from the one it got transmitted.

However, few investigators worked with semantic dependent model for resolving the issues in the IoT environment, that are following,

Sathyapriya et al., [Sat, 20a] presented a survey on IoT semantic technologies in healthcare. The author discussed the role of semantic concepts in IoT, IoT based healthcare domains and various ontology for handling IoT data.

Alhakbani et al., [Alh, 19] presented a system for managing IoT semantic data. The algorithms were proposed for matching events which utilized construction of tree. Also, the proposed system maintained the systematic relation between crucial applications. The proposed system performed better than the existing system but the process was not supported in the distributed environment.

Manonman et al., [Man, 20b], proposed the models for healthcare domain based on the semantic techniques. Initially, the processes of data gathering, data cleaning, and feature extracting and semantic modeling were performed. After that, IoT based medical care system was proposed that uses Health IoT ontology. The proposed system provided the better improvement in decision making.

Naveed et al., [Nav, 19], presented a model for resolving the problem of interoperability in IoT environment. This issue was happened during the conversion of medical data among organization. Two algorithms such as Word2Vec and Doc2Vec were utilized for matching related semantics in the dataset. The presented model eliminated these issues by transferring data into a regular form. But, the proposed model was only applicable for data of Electronic Health Record (EHR).

Various genres of these IoT Semantics Techniques will be discussed below.

2.3.1 Ontology for IoT

The term Ontology was found along with IoT since the works of [Bar, 12], [Kot, 12], [Jen, 17], [Ala, 17], [Shi,18]. Many works concentrated on solving the issues of improper interoperability and interpreting the ambiguous meaning of the data exchanged from one device to another including the conventional work [Kot, 12].

A comprehensive Ontology for IoT was studied and interpreted by [And, 18] to eliminate the difficulties caused due to the improper interoperability. Their study made use of more than 100 other studies including but not limited to [Goe, 17], [Jia,17] dealing with the same topic of Ontology for IoT to prevent improper interoperability. Using their study, the study was able to draw the following answers:

- The reason for deployment of the Semantics
- Futuristic themes available for IoT Semantic methodologies-eliminating interoperability
- Existence of various types of IoT Semantic methodologies-interoperability
- Extensive application of this interoperability maintaining IoT Semantic methodologies

Not just few works have concentrated on IoT Semantic methodologies, but there were many works proposed and implemented in order to sort out the interoperability issues. The concerned works are as follows:

- LIoPY - A legal compliant ontology to preserve privacy for the Internet of Things was proposed by [Lou, 18]

- Modified version of SSN - Semantic Sensor Network was suggested by [Hal, 17] with the help of ontology for IoT.
- An improvised system of IoT Semantic methodology was developed by [Gyr, 18a] in order to enable interoperability by the consideration of four IoT-relevant ontology catalogs.
- A developer-oriented methodology was suggested by [Baq, 19] in various diversified applications of Semantics and IoT.

[Gan, 17] also made a comprehensive comparative study of IoT Semantic methodologies aim to eliminate/ sort out the issues of interoperability. Not just this review was concerned about the proper interoperability by using IoT Semantic methodologies but various works have made their efforts in contributing to the same, are as follows: [Gyr, 18b], [Baj, 17], [Rah, 20], [Reb, 18], [Nag, 21], [You, 19]

Unless information retrieved from the IoT environmental systems were insightful and properly [Jia, 17a], the application of Semantics in IoT will be indispensable and be used in the future applications of IoT, which includes but not limited to Health care. Various levels of Interoperability as identified by [Har, 14] is indicated below.

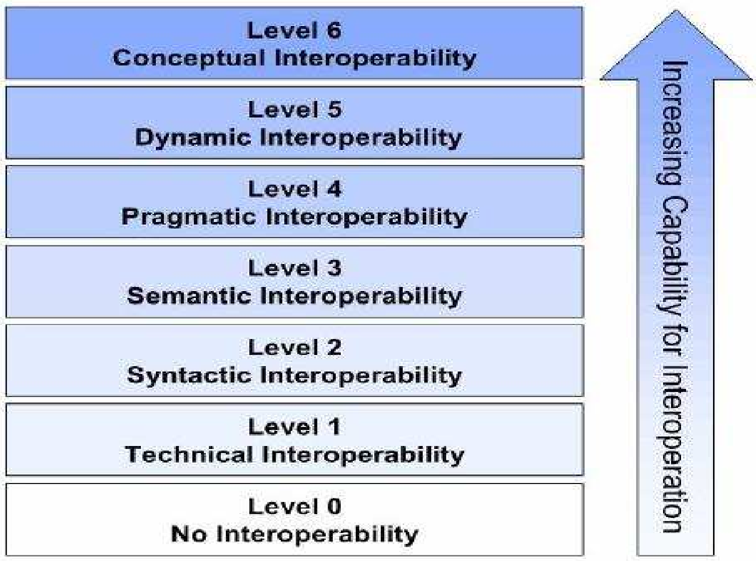


Fig. 2.2 Various levels of Interoperability [Har, 14]

Various levels of interoperability were exhibited in the fig. 2.2. Starting from no interoperability from level 0, the levels of interoperability improve to conceptual interoperability in level 6 at the modelling level based on the factor of increasing capability for interoperation.

2.3.2 Ontology for Healthcare

With many applications of semantic-oriented Ontologies in IoT increasing as the years go on, the need for applying it in the health care application increases as the health care development can't be overlooked in any means.

Thus, ontology is widely used in the arena of Health care. Some of the recent works are discussed one by one.

- A smart health care system for Senior Citizen people was discussed by [Gup, 21].
- A mined data-oriented SMO-Sequential Minimal Optimization with EAA-Enhanced Apriori Algorithm was suggested by [Sor, 20].
- [Kio, 19b] proposed a FHIR-Fast Healthcare Interoperability Resources of the Health Level 7 on the basis of base work [Kio, 19a].
- An integration of EHR systems in an IoT environment was achieved with the help Ontology Middleware [Ala, 18].
- COPD - Chronic Obstructive Pulmonary Disease was developed with the efforts of [Aja, 18] to support the Ubiquitous Healthcare Systems using Ontology-oriented model.
- [Che, 19] put forth a remotely powered diabetes diagnostic and controlling methodology using their ontology-based model.

- Wearable sensors were deployed by [Ali, 18a] to put forth a Fuzzy ontology and Type-2 Fuzzy Logic to improvise the capabilities of knowing the risks factors of various patients.
- A food habitat suggesting methodology was developed by [Sub, 19] to aid various patients.
- A system of NER-Named Entity Recognition was proposed by [Bat, 19] to detect and categorize various diversified diseases in the health care.
- SAREF- Smart Appliances REFerence, an industry specific standard-based methodology was suggested by [Mor, 18]

2.3.3 Analytical Survey

According to [Inf, 21], a typical Analytical Survey was the investigative studies that were taken up to analyze how any response variable like Disease existence could be related to any specific explanatory variable or variables. The analytical survey being carried out in any work depends upon the nature of application specifically.

Some of such analytical surveys taken up by various researchers in the domain of health care are presented below.

- An affected piece of brain was identified by [Moh, 21].
- A detailed analytical survey of various advanced computer technologies and IoT powered tools were done by [Rat, 19] to improvise the ambulance operating conditions.
- Cancer patients were aimed to be aided in their treatment by [Rah,19] with the help of a BlockChain and IoT-oriented multiple sensors operated system in home conditions.

- A smarter case study was made by [Ara, 17] to optimize the diabetes disease managing system.
- Another instant of analytical survey of various methods and classification methodology in IoT environment was done by [Bis, 20]

2.4 Deep learning for IoT

As IoT has already started as key element of the future Internet, many researchers have shown interest in experimenting the deep methodologies in an IoT environment [Ma, 19] surveyed about implementation of deep learning on IoT under the aegis of health care, smart industry, smart transportation, and smart home.

The dependency of third party software was studied by [Tan, 17] by transferring the TensorFlow, a deep learning-based platform into IoT environment.

A review was carried by [Ati, 20] to support and aid the smart city establishment by adopting IoT analytics and Deep learning methods.

A very useful system of computer aided cancer classification was developed by [Raj, 20] through using DNN and by employing fog computing concept. [Zhou, 20] devised an enhanced HAR (Human Activity Recognition) for the IoHT (Internet of Healthcare Things).

An innovative service model based on semantics was proposed by [Xie, 20] that permits an easier subscription by the end user to a certain physiological parameter amongst the existing sensed data. This model helps in achieving more effective healthcare system.

2.4.1 Uses of deep learning techniques in healthcare

Though many of the uses of deep learning techniques were reported in the earlier literatures, [Est, 19] made an effort to a devise comprehensive review of them in a more concise way by concentering on the health care. Many such reviews and researches were found as follows:

- [Shi, 17] put forth a comprehensive study of modern trends taking place in the deep learning methodologies taking into consideration the analysis of Electronic Health Record (EHR).
- Many clinical forecasting techniques of deep learning concept were studied and benchmarked by [Pur, 18] in order to forecast the conditions like CD-9 code group, hospitalization stay length, and mortality rate.
- A human oriented methodology of diagnosing the Diabetic Retinopathy was suggested by [Bee, 20] for deployment in the clinical conditions.

2.5 Importance of Diabetes disease prediction using IoT

Reducing the mortality in any hospital pertaining to a critical illness was always a tedious tasks, which doesn't exclude less-predictive disease like diabetes [The, 21]. Also, treating and reducing the risks of a patient admitted in an ICU unit was even more challenging, yet could be achieved with the help of deployment of proper deep-tech methodologies under IoT.

A 2-Dimensional CNN-Convoluted Neural Network Model was proposed and implemented by [Yil, 19] to develop an automated diabetes identifying system with the frequency spectrum images, extracted out of signals coming from the heart.

Having known the similarities between diabetes and cardiovascular diseases, [Ali, 17] made use of Bayesian Networks and ANNs- Artificial Neural Networks to predict them and avoid any critical illness conditions.

A decision tree and random forest-based machine learning methodology was proposed by [Zou, 18] to forecast the diabetes mellitus in order to avoid one in ten adults in the globe suffering from it. Another instant of decision supportive methodology was developed by [Yah, 19] with the help of both the machine learning and deep learning methodology to forecast diabetes.

A Home-oriented IoT diabetic forecasting was done by the efforts of [Bha, 20] by considering the inputs extracted from the urine. The data for the work was gathered by mounting the sensors across the considered toiletry. They made use of 4 components that aided to successfully classify, mine, and extract the gathered data to make decisions towards the effective forecasting of diabetes.

A smarter IoT environmental Diabetic disease identification was achieved by [Haq, 20] with successfully implementing the vital elements like data pre-processing, feature selecting, and diagnosing the diabetic condition. The feature selection was executed by Random Forest and Ada Boost along with the filter operation executing Iterative Dichotomiser 3. The Kaggle machine learning data repository was used to input for the modelling and testing phases of the work.

Likewise, [Sah, 20] made an extensive performance evaluation of health decision making system by taking into consider both the machine learning and deep learning CNN methods.

Not only taking care of the critical illness is important, but also avoiding those critically illnesses by predicting of Diabetes disease also becomes important and significant owing to the current raise of diabetes patients all over the globe.

2.6 Issues and Challenges

Various challenges and issues pertaining to the adoption of E-Health in IoT was investigated and surveyed by [Sca, 17].

Areas of difficulties realized while applying any deep learning methodology in an IoT environment as per [Ma, 19] were as follows:

- Data Collection
- Hardware configurations
- Model Training
- Design of the system

Avoiding heterogeneity and interoperability will also be another concern when investigating the well-optimized data gathered. Slower reaction times in health care application like diabetes disease prediction cannot be compromised.

Thus, proper and timely detection of the diseases should have to be made by developing an efficient classifying methodology.

2.7 Chapter Summary

The chapter had discussed various works found in the literature pertaining to the Diabetes Disease Prediction and IoT / Ontology / Semantic oriented methodologies by categorizing into various sub-sections with reference to all the basic concepts and terminologies used.

Chapter – 3

Methodology for Diabetes Disease Prediction

CHAPTER - 3

METHODOLOGY FOR DIABETES DISEASE PREDICTION

3.1 Introduction

This chapter describes the research methodology of the proposed models. The research objectives are explained clearly in this chapter and then the overall process of the research work is depicted using a methodology diagram. This chapter also elucidates the implementation system, dataset collection, software, and hardware packages deployed in the whole research process for the diabetes disease prediction.

3.2 Research Objectives

The primary purpose of the research is to propose a diabetes disease prediction model with improved accuracy for the IoT environment using the following models:

3.2.1 Rule Embedded Ontology-based DNN Model

The ontology-based model is used for data preparation, in which the data has been collected from various sources or devices and to classify the diabetes disease prediction [Sat, 20b]. The domain-based ontology is framed to enrich the data and various rules are then constructed for validating that ontology. The deep neural network is constructed for diagnosing the disease using the ontology data, where the linear activation function is utilized to activate the results. The ontology data is classified into either diabetes or non-diabetes types based on the DNN model.

3.2.2 Deep Learning with clustering and optimization functions based model

Deep Learning with the clustering and optimization function based model is used to predict the risk of diabetes disease. The classified and enriched data is given as an input to this model. The input of the network passes the information to the next hidden layer where K-means clustering algorithm is applied for grouping the common items based on the Euclidean distance to find out the risk for any particular person. Additionally, binary multi-objective optimization is utilized for improving the prediction result. Finally, output layer uses the Softmax function as an activation function to produce the predicted output. The result is predicted as low risk or high risk or medium risk cases of diabetes. The performance of the model is then being compared with other machine learning and deep learning models for validation.

3.2.3 An enhanced deep learning model

The enhanced deep learning model is proposed to enhance the accuracy of the above prediction model. This is accomplished by combining the XGBoost algorithm and Adam optimization with the preceded model. The XGBoost algorithm is optimized and then embedded into the model which predicts the disease accurately than the previous model. It also, attains the requirements which were not satisfied by the previous model.

In the next section, a methodology diagram is drawn and indicated to attain the projected research objectives, which have been explained in the succeeding section. The proposed models are demonstrated and presented with simulation results as and when needed along with the tools utilized for simulation.

3.3 Methodology Diagram

The methodology diagram contains two important phases, namely the data collection phase and the analytics phase. In the data collection phase, data is collected through various modes like sensor mode, questionnaire mode, and clinical report. Sensors play a vital role in this data collection, which usually many of the data needed for the whole work. The personal information is being collected through the questionnaire and the medical history of any individual is being collected from various clinical reports. In the data analytics phase, three prediction models are proposed, namely: RESOC_DNN, IMO_DNN, and EIMOXB_DNN.

Various IoT sensors and deployed questionnaire patterns in this research work are explained with the help of methodology diagram in the below fig. 3.1. The projected research objectives are achieved by adopting the depicted methodology.

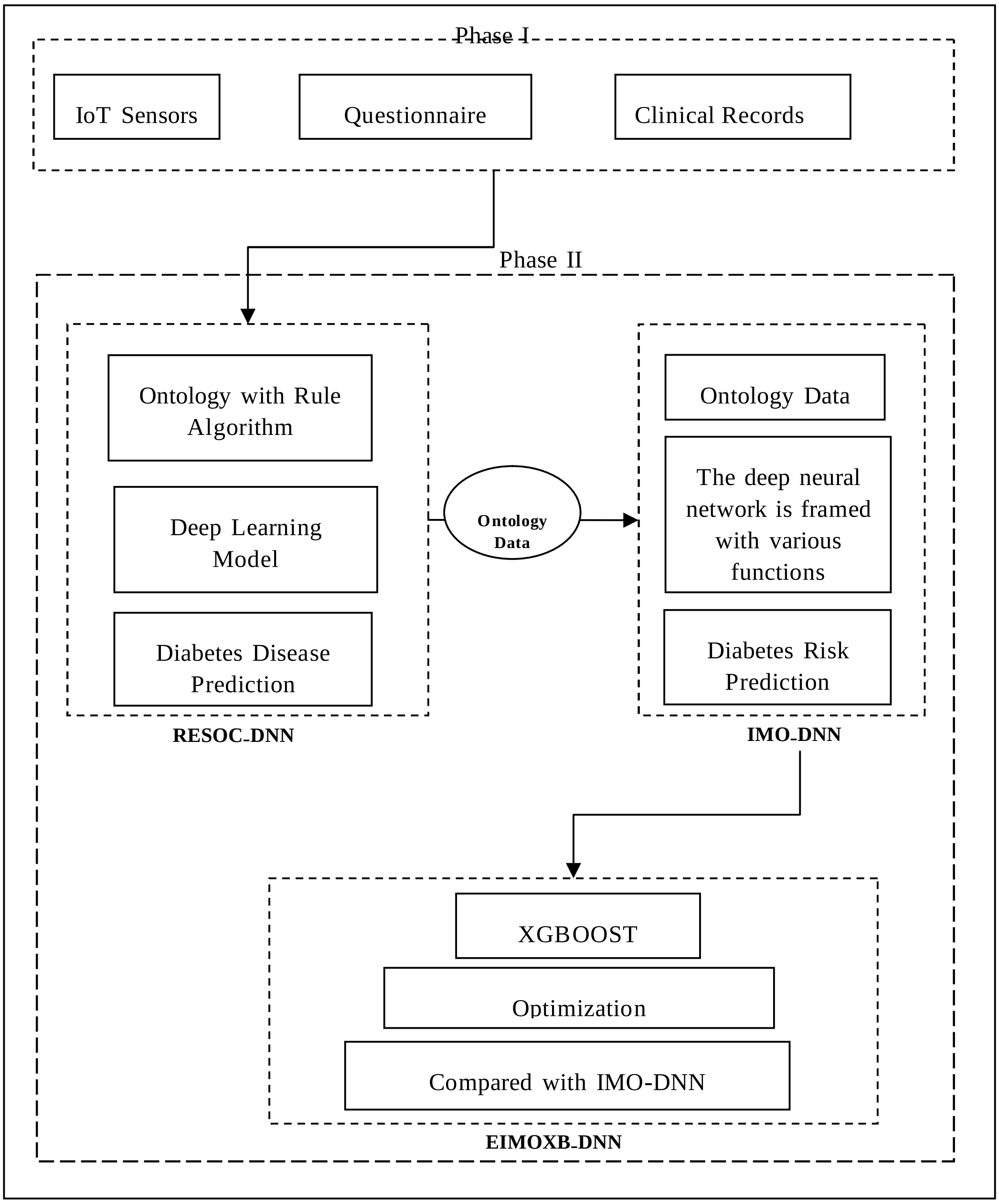


Fig. 3.1 Methodology Diagram of IoT Semantic Based Deep learning models for diabetes disease prediction

3.3.1 Phase 1: Data Preparation

In the Phase 1, data are collected through various sources such as IoT devices, questionnaire forms, and clinical records. Thus, there are three types of data involved, namely: sensor data, clinical data, and questionnaire data [Aro, 19a].

Various sensors are used for monitoring and collecting numerous health-related data of patients. The questionnaire form contains various questions that include the details about age, gender, food habits, heredity disease, height, weight, and medicine intake, etc., Additional data are collected through clinical records that include medical history of the patient or any current medications that are being taken at present, etc.

I. Sensor Data

IoT devices are attached to monitor and track the patient health status. The five types of sensors used are as follows: heart rate sensor, temperature sensor, step count sensor, blood pressure sensors, and blood glucose sensor. These devices continuously generate data about the health status of patients, which are being collected and stored in the cloud. The detailed explanation of all these sensors are discussed in the following section,

a. Heart rate

A heart rate sensor is a personalized monitoring device that allows to measure and record the heart rate. It is often used to collect heart rate data while performing various types of exercises. The normal heart rate for healthy individuals is 60 to 100 beats per minute. The average heart rate is 70 bpm in adult males and 75 bpm in adult females [Pal, 02].

b. Temperature Sensor

Human body temperature is simply determined by body temperature and the amount of heat radiated by the body. A normal person's body temperature depends on different factors such as ambient temperature, the person's gender, and eating habits.

In healthy adults, it is between 97.8°F (36.5°C) and 99°F (37.2°C) [Gen, 19]. Various other features like fever, low body temperature or any other illness that can lead to a change in body temperature are also considered. Also, fever is a common indicator of almost all diseases, thus, it needs to be handled with proper care.

c. Blood Pressure (BP) sensor

Blood pressure (BP) is considered to be the most important cardiopulmonary parameter, which refers to the pressure exerted by the blood against the arterial wall. BP gives indirect information about blood flow during the contract (systole) and rest (diastole), which may also indicate cellular oxygen delivery. Ambulatory BP monitoring allows taking BP measurements several times a day, which is ideal for monitoring high blood pressure (hypertension). This improves the prognosis of global morbidity and cardiovascular disease [Gro, 17].

d. Blood Glucose Sensor

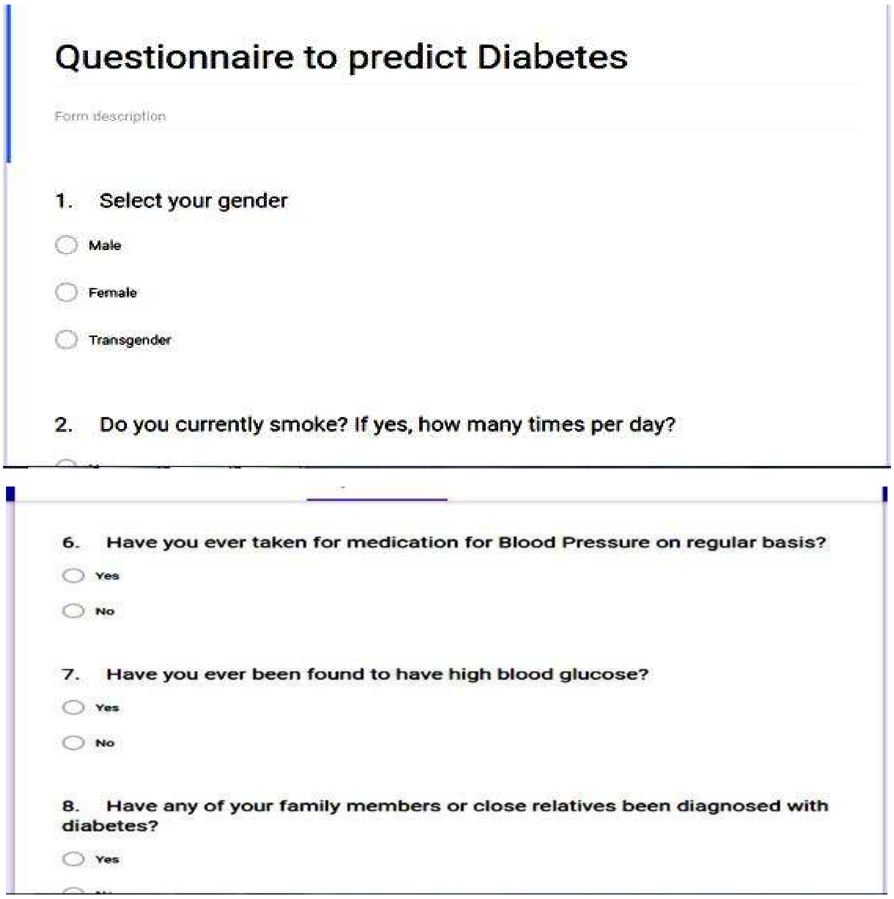
Glucose sensors are used to measure a patient's blood glucose concentration and are an important part of managing diabetes. Normal blood glucose levels range from 80-120 mg / dL and it can group to 250 mg/dL after meal. These sensors are usually capable of measuring peak blood sugar levels (between 20-500 mg / dL or 1-30 mM) that a patient may experience during hyper or hypoglycaemia and contains a resolution of ~ 1 mg / dL or ~ 50 µM [Rag, 19].

e. Step Count Sensor

The step counter sensor is used to get the total number of steps walked by the user. A pedometer is usually a portable and electronic or electro mechanical device that measures every step a person takes by detecting the movement of a person's arm or hip. Because the distance of each step taken by an individual varies, an informal calibration is required to be defined by the user by providing the covered distance in a unit length (such as kilometres or miles), although the pedometer uses electronics and software that automatically determines how the person's steps vary. It is mostly used in sports and physical fitness centers. It is also used in healthcare for monitoring various physical activities of each patient.

II. Questionnaire

The questionnaire is prepared by using multiple-choice questions to collect personal as well as the health details of various patients. This questionnaire is prepared based on patient age, gender, habits that spoil their health like: smoking and alcohol drinking, height, weight, medication taken by an individual, family history, work type and physical activity. The model questionnaire is shown in fig. 3.2.


Questionnaire to predict Diabetes

Form description

1. Select your gender

- Male
- Female
- Transgender

2. Do you currently smoke? If yes, how many times per day?

6. Have you ever taken for medication for Blood Pressure on regular basis?

- Yes
- No

7. Have you ever been found to have high blood glucose?

- Yes
- No

8. Have any of your family members or close relatives been diagnosed with diabetes?

- Yes

Fig. 3.2 Model Questionnaire to collect personal information

III. Clinical data

Clinical data refers to the patients who are already under any kind of treatment.

Table 3.1 exhibits the attributes of collecting data. The field column indicates the properties of the collected clinical data. These attributes are extracted separately and framed in a unique format.

Table 3.1 Dataset and its descriptions

Field	**Data Type**	**Description**
Gender	Categorical	It includes the options like male, female or transgender
Age	Numeric	This column represents the total number of years completed by the person since they're born.
Glucose	Numeric	This column represents the blood glucose value of the person
Blood Pressure	Numeric	Indicates how much pressure there is in the blood against the artery walls when the heartbeats
Skin thickness	Numeric	It defines the thickness of the skin.
Insulin	Numeric	It defines the insulin level of the human body
BMI	Numeric	Calculating the weight(in kg) and height(in m) of a person as Body Mass Index
Pedigree function	Numeric	Assessment of the likelihood of diabetes based on family history
waistline	Numeric	Represents the measurement around a person's body at the waist
Calorie intake	Numeric	Refers to daily caloric intake
Physical Activity	Text	It defines the regular exercises of a particular person.
Medicine intake	Text	It defines if the person already takes any medicine or not.
Food habit	Text	It defines which type of food they eat
Pregnancy	Numeric	Number of times pregnancy
General Habit	Text	Habits that spoil their health like: smoking and alcohol drinking

3.3.2 Phase 2: Disease Prediction

Usually, health related data is gathered from online repositories. But the data collected in real-time using smart devices and other sources has some drawbacks such as heterogeneity and data formats, etc. The heterogeneity problem is faced when the data was collected from various IoT devices or various sources like questionnaires, clinical reports, etc.

These problems may reduce the performance of the prediction models, thereby increasing the complexity of data and the processing time. The semantic-based concept gives a solution for these heterogeneity and data formats. Various rules are being created to improve the performance of the model.

This phase presents three models for diagnosing the diabetes disease accurately by avoiding heterogeneity and maintaining a unique format. The following section explains the proposed models.

A. Rule Embedded Semantic Ontology-Based DNN Classifier Model (RESOC_DNN)

The proposed model contains two phases like semantic development phase and the prediction phase. The collected details are sent to the semantic module, where the data is converted into a meaningful and understandable format for quick analysis. The quality of data is very important to improve the analysis results. Various data sources are handled by ontology, which deals with the enrichment of data to combine all data for making concepts and relations in order to store them as a single format.

Ontology represents the described knowledge expressed as a triple form, namely: subject, predicate, and object by using an RDF graph. Subject denotes resources while the object denotes values and the predicate indicates the properties or features of the properties. This reveals the relationship between subject and object. At first, the basic terms and concepts are specified and then grouped into classes, subclasses, object properties, and data properties. After that, a set of rules are embedded into the ontology for handling real-time data from various sources and validating them. Then, the deep neural network model is formed with input layer, hidden layers, and output layer. This prepared ontology data is sent to the model and the network is trained based on it. After training, the test data is sent to the network and the result is predicted using the activation function. The accurate result is gathered from the network and the performance of the model is compared with other models based on the evaluation metrics as applicable.

Ontology is implemented using the protege tool and the proposed model is deployed in Anaconda tool using Python language over the windows operating system. Finally, a confusion matrix is being deployed to evaluate the performance of the model. The performance is evaluated using the metrics of precision, recall, accuracy, and f - score.

The proposed RESOC_DNN model predicts the diabetes disease accurately. Moreover, finding the risk level of diabetes assists various patients in protecting them from a critical severity and also reduces the workload of doctors. But this model is found inappropriate for predicting the risk of diabetes. So, for further enhancement of the model and to predict the risk level of diabetes, a second model is being proposed.

B. IMO_DNN model for predicting the risk of diabetes patients

This proposed model IMO_DNN follows the structure of conventional deep learning technique that contains input layer, hidden layer, and output layer. The Diabetes ontology data is given to the input layer as an input and the result are predicted by trained deep layers. After the prediction of diabetes disease, the data is given to the IMO_DNN model, where data is unsupervised because there is no class label. The proposed model is constructed to predict the class label into the following: low risk, medium risk, and high risk. In the hidden layer, K means algorithm is utilized for clustering data items using similarity measure to predict the class. The softmax activation function is used in the output layer that predicts accurate results based on deep trained layers. Then, the model is improved by binary multi objective optimization function to reduce the loss function and improves the performance of the model.

The IMO_DNN model has predicted the risk of diabetes disease. The proposed model is evaluated for its performance by the metrics like accuracy, precision, recall and F1-score with other traditional models. The proposed IMO_DNN model was found to have problems with gradient disappearance and gradient explosion. Also, it's learning ability was limited, which often failed to achieve the desired results in practical tasks. Therefore, an improvised model is proposed which makes some improvements to the structure of memory.

C. Enhanced IMO_ DNN with Extreme Gradient Boosting (EIMOXB_DNN)

XGBoost could be defined as an additional model in the output layer to improve the accuracy of risk prediction. The proposed EIMOXB_DNN model enhances the IMO_DNN model by handling the gradient disappearance and thus results in improved accuracy in predicting the target value. The EIMOXB_DNN is enhanced by Extreme Gradient Boosting and optimized by Adam optimizer.

The enhanced EIMOXB_DNN model provides optimal decisions in predicting the risk of diabetes. The accuracy of the EIMOXB_DNN model produced high accuracy than the IMO_DNN model.

3.4 Evaluation Metrics

There are numerous evaluation metrics to measure the accuracy of predictive analysis. The most commonly used assessment metrics are: accuracy, recall, F-score, and precision, which are used to determine the accuracy of the proposed models. Common terms used for evaluation measurements are:

- *True positive (TP)* (properly recognized)
- *True negative (TN)* (properly discarded)
- *False positive (FP)* (improperly recognized)
- *False negative (FN)* (improperly discarded)

3.4.1 Confusion Matrix [Moh, 12]

It is a table used to describe the performance (precision, recall, f score and accuracy) of a prediction model or classification model on a set of experimental data. It is also called the error matrix that is represented in table 3.2.

Table 3.2 Error Matrix or Confusion Matrix

Characteristics	Positives	Negatives
Actually positive	Amount of True positive (TP)	Amount of True Negative (TN)
Actually negative	Amount of False positive (FP)	Amount of False Negative (FN)

A. Precision

Precision is defined as the 'quality of being accurate' and refers to how close two or more measurements are close to each other, regardless of whether those measurements are accurate or not. The equation for computing precision is given in eq. 3.1.

$$Prec = \frac{TP}{TP + FP} \quad \text{... Eq. (3.1)}$$

B. Recall

This is a fraction of the documentation associated with the query successfully recovered. The equation for calculating the recall is given in eq. 3.2.

$$Rec = \frac{TP}{TP + FN} \quad \text{... Eq. (3.2)}$$

C. F-score

It is a metric that joints precision and recall. The average of precision and recall are given in the traditional F-score 3.3 equation.

$$F = 2 \cdot \frac{Prec.Rec}{Pre + Rec} \quad \text{... Eq. (3.3)}$$

D. Accuracy

The accuracy of predictive analysis is calculated using the subsequent equation. It is given in eq. 3.4.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{... Eq (3.4)}$$

3.5 Dataset and Software Packages

3.5.1 Dataset

Data collection is done in various ways. The total number of instances 1359 and the number of attribute is 14. The table 3.2 exhibits the description of the datasets.

3.5.2 Software Packages

Choosing the best tool for IoT based clinical data analysis can be a daunting task for entrepreneurs and academics. This section provides a comprehensive overview on software packages that used for the models which proposed for IoT semantic based Deep learning (DL) models to predict the diabetes disease.

A. Protege [ElA, 18]

Protege is a powerful editing environment with many features. This is a quick tool to create our standard ontology in quick periodic time. It provides a GUI to define ontology. It also includes the excluded classifiers to ensure that the samples are consistent and to infer new information based on the analysis of the ontology. Protege, like an eclipse, is a framework in which various programs recommend plug-ins. This application is written in Java and uses the Swing function extensively to create user interfaces. The health ontology is developed using Protege, which combines three forms of the dataset into a single ontology. XML is the coding used to embed the data into the tool and the combined data is then stored in RDF format. RDF is a mark-up language that enables the exchange of highly interconnected data. Its three-part structure allows for the identification and disambiguating of resources. Finally, the data is interpreted in JSON format for analytical purposes. Fig 3.3 illustrates the start-up process of protege with sample feature configuration.

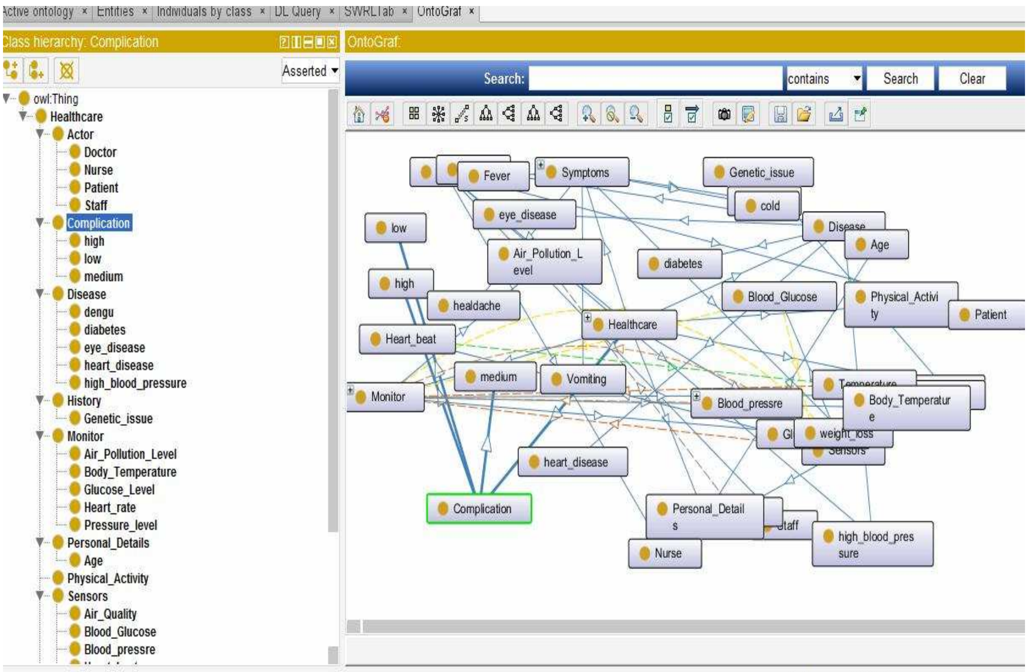


Fig. 3.3 Prepared Ontology for IoT Based Diabetes data

B. Arduino [Pat, 17]

The Arduino Integrated Development Environment (IDE) is a cross-platform application written with functions from C and C ++. It is used to write and upload programs on Arduino compatible boards. It is also used with third-party cores with the help of other vendor development boards. The source code for the IDE has been released under the GNU Public License, Version 2. The Arduino IDE supports both C and C++ languages using special rules of the code structure. The Arduino IDE provides a software library from the wiring program that provides many common input and output procedures. User-written code requires only two basic functions: sketch and start the main program cycle, a program compiled with Stub Main (), integrated with the GNU management program and included with the IDE distribution. The ArduinoIDE is successfully installed and coding are framed to receive and send data from sensors to the cloud.

C. Thingspeak [Par, 19]

Thingspeak is an open-source Internet of Things (IoT) application that stores and retrieves data over the Internet or a local area network using the API, HTTP, and MQTT protocol. Thingspeak allows creating a social network with recording apps, sensor applications, location tracking apps, and status updates. Thingspeak has been integrated with support for MathWorks from math computer software - MATLAB. So, users do not need to purchase a Matlab license to analyze and display uploaded data.

The user account is created first with a valid username and password. Thingspeak provides eight fields for the channel of each account. It offers both private channels and public channels. The Channel settings include: channel name, description, location, URL, video, and tags. Additionally, API keys are automatically generated to read and write sensor data as soon as the channel gets created. Stored data can be downloaded in the form of .csv or .xml or .json and other datasets from some other sources can also be imported. The fig 3.4 shows the created channel in thingspeak to collect sensor data.

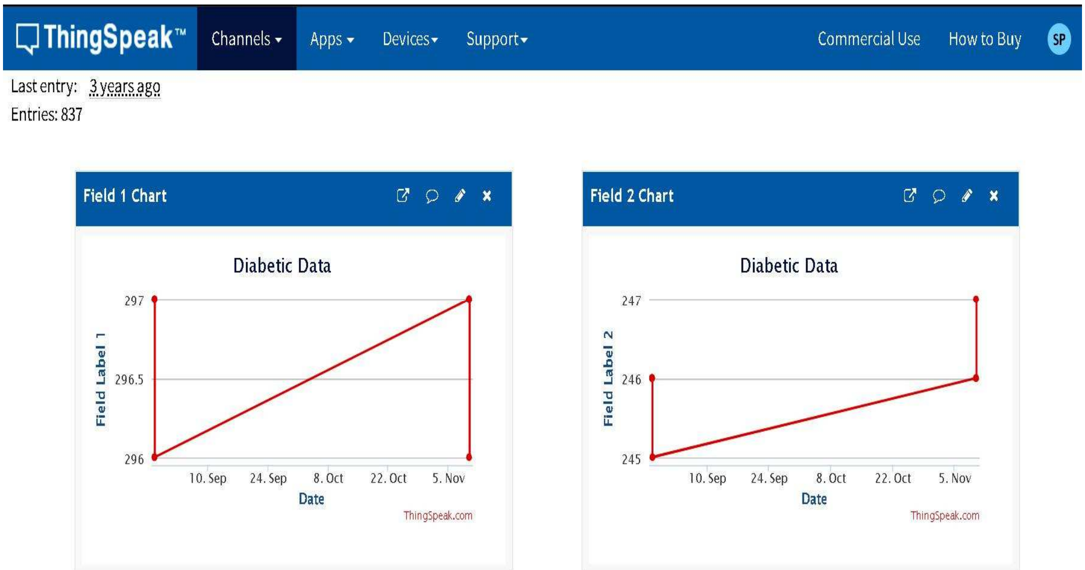


Fig. 3.4 Data Storage in Thinkspeak

D. Anaconda [Pat, 21]

The Anaconda is the distribution of Python and R programming languages for the scientific computer (data science, machine learning applications, large-scale data processing, and forecast analysis, etc.). It aims to facilitate package management and deployment. The distribution includes the data-science packages suitable for Windows, Linux, and MacOS. It got developed and maintained by Anaconda, Inc., which was founded in 2012 by Peter Wang and Travis Oliphant. It is also known as the Anaconda Distribution or Anaconda Individual Edition, while the other products of the company are the Anaconda Team Edition and the Anaconda Enterprise Edition, both of which are not free.

Package versions in Anaconda are maintained by the package management system-Conda. This package manager was released as a separate open-source package because it was useful for anything other than Python. There is also a smaller, bootstrap version of Anaconda called the Miniconda, which includes only Conda, Python, their dependent packages, and a limited number of other packages.

3.6 Chapter Summary

In the above sections, three main objectives of the research work were briefly explained. The methodology diagram given in this chapter gave a crystal picture of the proposed work. Datasets used for this work were displayed via sensors, questionnaires, and clinical reports. The overall methodology of IoT semantic based deep learning models for diabetes disease prediction was discussed. This was carried out based on three models, namely: RESOC Model, IMO-DNN Model and EIMOXB_DNN model. The mentioned models are analyzed and also discussed in the following chapters where the evaluation metrics will be used to measure the performance of each model.

Chapter – 4

Rule Embedded Semantic Ontology-Based DNN Classifier Model (RESOC_DNN)

CHAPTER - 4

RULE EMBEDDED SEMANTIC ONTOLOGY-BASED DNN CLASSIFIER MODEL (RESOC_DNN)

4.1 Introduction

In the IoT data analysis, the semantic-based approach plays a vital role in handling the heterogeneity issue among other domain-based datasets. The data are collected in real-time from various sources and diverse devices for the prediction of diabetes disease. This chapter explains the ontology preparation, rule generation, model creation, and disease classification for the medical data which have been collected from various sources.

In general, the data collected from various sources or sensors may have various formats. This may cause problems in data exchanging and task analyzing. Moreover, enriching data and inferring knowledge on time with favorable accuracy for a particular domain are other daunting tasks. So, the proposed model mainly concentrates on ontology preparation, reasoning rules generation, and disease classification. The proposed RESOC_DNN Model handles the heterogeneity in data and classifies the diabetes disease with an acceptable level of accuracy.

4.2 Background

The background behind the proposed method of diabetes diseases prediction is discussed in the succeeding sections.

4.2.1 Ontology

Ontology is the part of the Semantic Web that describes the concepts and the relationships between every collected piece of information using systematic profiles for avoiding heterogeneity issues [Xia, 18, Dud, 18]. It provides a simple and great idea to categorize a concept into a superclass and subclass and further, it acts as a framework for describing the knowledge by categorizing and interpreting the information [Asi, 18]. In addition, ontology also allows the analysis and possible reuse of the obtained knowledge. Thus, ontology enables the sharing of categorized information across people and applications as if it was a general vocabulary. Thus, it results in the development of infrastructure, which could favor the exchange of information and innovation.

4.2.2 Reasoning with rules

The reasoning is the process by which automated events could create new associations based on data with the help of some extra information in vocabulary form [Dis, 20]. In the semantic web, such extra meaning could be added by a set of rules or vocabularies [Hoh, 20]. Logical reasoning ensures and resolves the conflicting knowledge due to deficient notice.

4.2.3 Deep Neural Network

Recent studies have shown that the deep neural network (DNN) could be a great solution to more complex classification problems [Bas, 19]. DNN based classifiers are very effective because the accuracy of the classification ratio is very high in contrast to the conventional classifiers. It attempts to create hidden patterns from the source data set. Deployment of these hidden forms is better than using raw features, which could enhance the DNN classification performance [Liu, 17]. However, there were only a few studies in IoT-based disease diagnostic systems dealing with these technologies. Some researchers have recently proposed diagnostic models based on

Deep learning. However, these models do not provide the expected outcome in the way that they aren't applicable in real-time. So, there is a need to improve the performance of the diseases diagnostic model [Che, 17], [Cal, 17].

4.3 Related works

Minoo et al., [Min, 20], presented the semantic model with a set of rules for data assimilation, standardization, and dimensionality reduction to tackle the problems in time consumption and difficulty in handling the heterogeneous data source. The proposed approach provided standardized medical data by avoiding heterogeneity. The author made use of the machine learning features in their approach to reduce the dimensionalities of data. However, it was not suitable for selecting the best subset as it could increase the computational time while working with semantics.

Balakrishna et al., [Bal, 20], made an investigation for various data integration and data analysis using machine learning algorithms for the IoT healthcare domain. Furthermore, various semantic and machine learning techniques for data integration were reviewed. Finally, future directions were discussed in the field of data integration taking place with the sensors in healthcare using semantic and machine learning approaches. However, their investigation for the healthcare domain was not implemented and validated using any tools.

Elsaleh [Els, 19], presented a lightweight IoT stream ontology for annotating streaming data. The model had been developed by following the most recognized guidelines of the semantic model and IoT environment. The well-known Semantic Sensor Network ontology for sensor descriptions was used in this developed lightweight model. The annotated data were extracted in RDF Triple format and finally, a few use cases, tools, and applications were discussed. Scalability and quick processing were the essential parameters that could have been improved in this work.

Chen [Che, 19], developed an ontology-based model for diagnosing, monitoring, and giving treatment to diabetes patients in a remote manner. The proposed ontology model solved the inconsistency problem by analyzing the patient information in detail. The performance of the proposed model was validated using Semantic Web Rule Language (SWRL) rules. Moreover, the experimental results proved that the model predicted the diabetes disease and recommended appropriate prescriptions. Anyway, the model was not suitable for any critical situation, because it was very time-consuming.

4.4 Objective

After going through the works reported in the literature, it could be inferred that numerous of the earlier works were done based out of ontology, however, none of the works has been carried out till now on the domain ontology, in particular by considering the data taken from multiple sources and analytical models that could in turn derive insights from ontology. An analytical model was not built to find out accurate insights. This has motivated the idea to build a new model for handling IoT based domain data and extracting information to make an efficient diseases prediction model. The objectives of the proposed approach are as follows:

- To develop an ontology that supports both IoT and domain data
- To develop a rule engine that generates rules for validating the ontology
- To develop a deep neural network model for disease classification

4.5 Methodology

The overall process of the proposed data analysis model has been divided into 3 phases. The following fig 4.1 explains the processing diagram of those phases as described below.

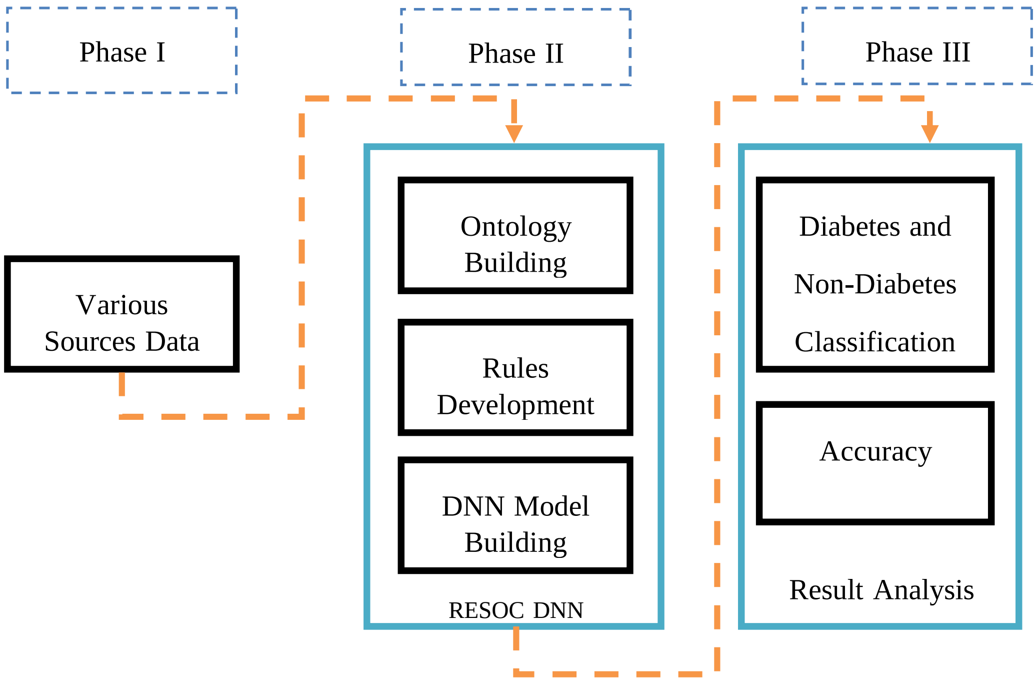


Fig. 4.1 Process flow of proposed RESOC_DNN model

4.5.1 Phase 1: Various Sources of Data

The data collection process has been described in the preceding chapter. The collected data are prepared in a required format by proposing a semantic algorithm. The semantic algorithm includes the process of ontology modelling and rule development. Various processes involved in this phase will be briefed in the upcoming subsections.

Table 4.1 shows the collected raw data. The raw data was gathered from various sensors, questionnaire, and clinical records that depicts in fig. 4.2. It consists of health matrices assessing features like age, BMI, temperature, blood pressure, blood glucose, pulse rate, food habits, physical activity, and calorie intake, etc. The proposed model provides a semantic annotation to these data to handle heterogeneity and provides better understanding through ontology. This prepared semantic data stored in the graph database will reduce space consumption and data loss. By generating various rules, a strong association is created among data in the ontology.

Table 4.1 Descriptions of Collected dataset

Field	Description
Gender	It includes the options like male, female or transgender
Age	This column represents the total number of years completed by the person since they're born
Glucose	This column represents the blood glucose level of the person
Blood Pressure	Indicates the pressure exerted by blood against the artery walls when the heartbeats
Skin thickness	It defines the thickness of the skin
Insulin	It defines the insulin level of the human body
BMI	Calculating the weight (in kg) and height (in m) of a person is Body Mass Index
Pedigree function	Assessment of the likelihood of diabetes based on family history
Waistline	Represents the measurement around a person's waist
Calorie intake	Refers to daily caloric intake
Physical Activity	It defines the regular exercises of a particular person.
Medicine intake	It defines, the person’s medications taken if any
Food habit	It defines which type of food they eat
Pregnancy	Number of times the person has been pregnant

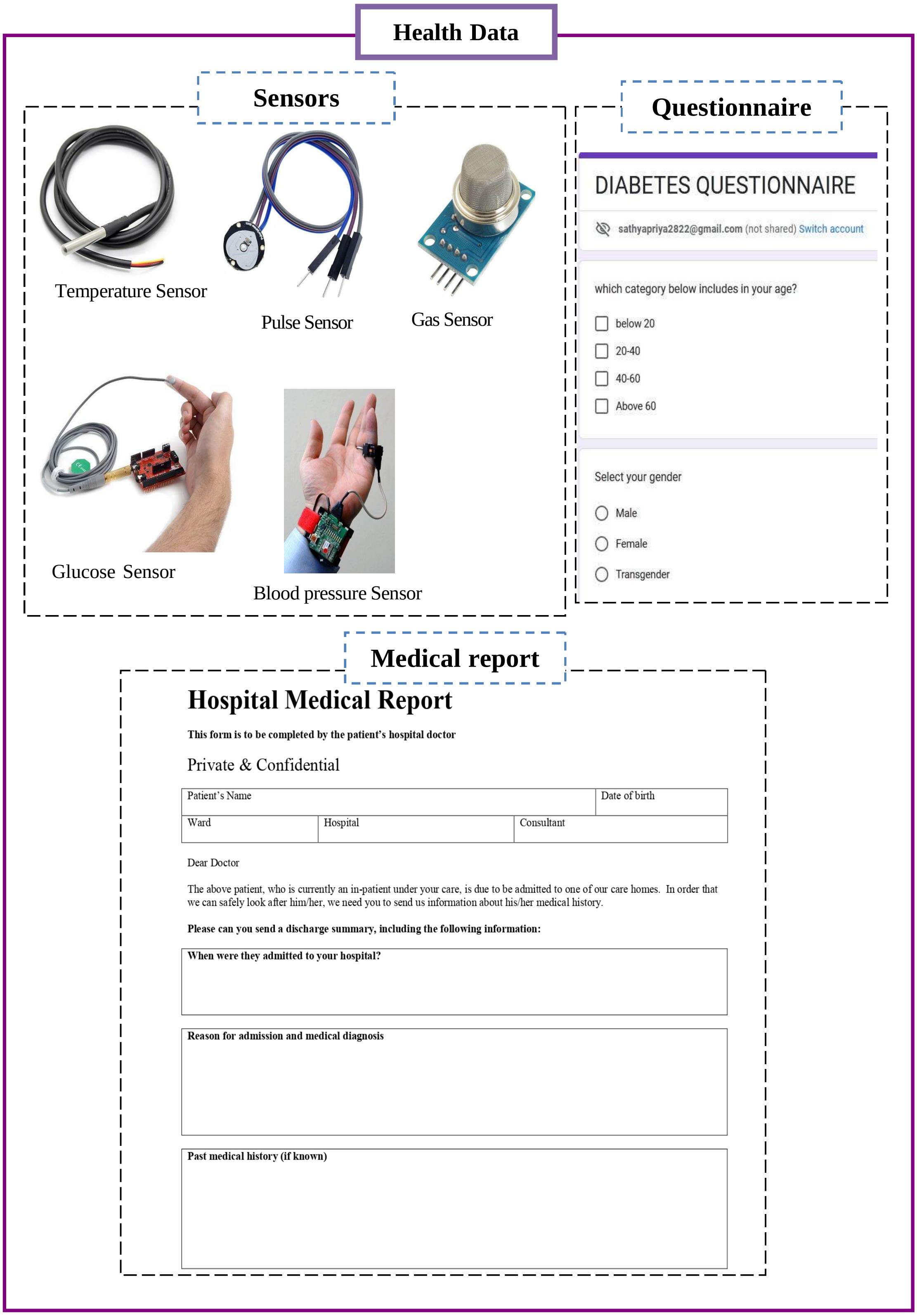


Fig. 4.2 Data collection from various sources

4.5.2 Phase 2: RESOC-DNN Model

The RESOC-DNN Model has been developed [Sat, 20b] for handling data from various sources and devices to perform analysis over it. For which, three types of algorithms are proposed viz: ontology algorithm, Rule generation algorithm, and DNN algorithm (For training and testing ontology data).

The model is proposed with different functions like ontology preparation, rule generation, and disease classification to have a thorough knowledge in classifying the diabetes disease in the healthcare domain. As a result, the prepared data will display the label into two as: "diabetic" and "non-diabetic". In the existing works, ontology was developed for IoT devices or common domains (healthcare, agriculture, etc.,) and rules were developed for evaluating the ontology. However, there was no proper analytical model using ontology data with these features. But in the proposed model, ontology is being prepared for a particular disease and various rules are then generated for inferring knowledge from it and also ontology is being validated. Furthermore, a deep learning-based model is developed to classify the disease accurately.

A. Ontology preparation

The raw data collected from various sources are found to be heterogeneous in nature, due to their formats and device types. The quality of data is very important to improve the analysis of results. Thus, the quality of data is being verified by ontology. The process deals with the enrichment of data, which is being achieved by combining all the data; developing concepts; and interpreting the relationships among them. It is then stored as an RDF graph in a fixed format.

Ontology represents the described knowledge in a triple form using an RDF graph, with parameters subject, predicate, and object. Subject denotes resources; object denotes values; and predicate indicates the properties or features of the resources, which reveals the relationship between subject and object. First of all, the basic terms and concepts are specified and then grouped into classes, subclasses, object properties, and data properties. Some of the terminologies of the proposed ontology are listed in the below table 4.2. Here, actors are subjects; while the relations are predicates; and the concepts are objects.

Table 4.2 Terminologies for Ontology

Actors	Concepts	Relation
Doctor, Patient, Nurse, Physician, Staff, sensors, Admin, and Manager	Temperature, Fever, treatment, blood pressure, Cardiac arrest, raised sugar, disease, eye ailment, Blood Pressure, Blood Glucose, and air quality	Has, treated by, affected by, has symptoms, has tested, has risk, has value, has side effects, etc.,

The procedure for preparing ontology is given as follows:

Step-1: Select raw data

Step-2: Prepare concepts for each feature

Step-3: Add properties to each concept

Step-4: The relations are identified among various concepts

Step-5: Annotations are made to the already added properties

Step-6: The graph of prepared ontology is generated

The prepared ontology for health data is shown in fig 4.3.

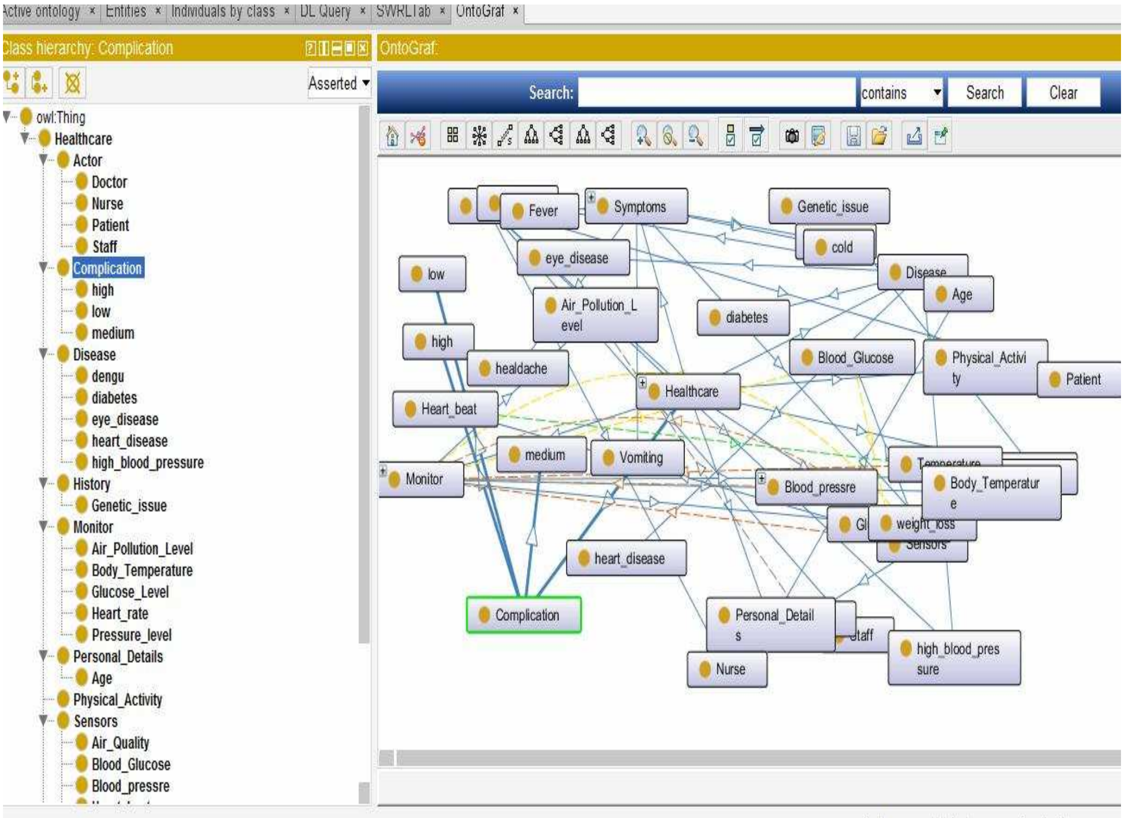


Fig. 4.3 Graph of the developed Ontology

Then, the proposed algorithms are discussed below.

Algorithm 1: Ontology Preparation

Input: Data from heterogeneous sources

Output: Semantically enriched Data

Define S = $\{s_1, s_2, \ldots s_n\}$ // **S means Sensor Data**

Define Q = $\{q_1, q_2, \ldots q_n\}$ // **Q means Questionnaire data**

Define Cl = $\{cl_1, cl_2, \ldots cl_n\}$ // **Cl means Clinical Data**

Define D = {S, Q, Cl} // **set of all collected data**

Define P, R, C

$c_i \in C$ // **Set of classes, where i = (1,2,3,...,n)**

$p_i \in P$ // **Set of properties, where i = (1,2,3,...,n)**

$r_i \in R$ // **Set of relations, where i = (1,2,3,...,n)**

```
Define O = {P, R, C}          // Ontology representation of collected Data D

Define G = {O}                // Graphical representation of Ontology O

    di ∈ D where i = (1,2,3,...,n) // Each data read from database D

    For (i=0; i<di; i++)

        {

        di −−> O;   // added ontology o

        if (di added in O)

          {

                visualize O;

           }

        else

           {

        try again

            }

        }
```

B. Rule Generation

The data preparation and categorization of the class label for domain-dependent action is a tedious task. Also, effective storage spaces are needed to store without any loss in data.

Numerous rules are embedded into the ontology for handling real-time data from various sources and better classification of results. The development of rules lies in discovering automatically the new facts based on an amalgamation of data. These rules create automated procedures for semantic triples that create new relationships from existing triples. Also, it plays an important role in detecting discrepancies in the integrated data. Here, the rules are like logic programs, which focus on finding the common relationships and defining common mechanisms for building those relationships based on existing ones. This proposed system optimizes the reasoning issues within data. The rules for inferring knowledge within the ontology are as follows,

Rule set 1:

1. Patient (? , x) ^ has_sensor (?x, temp) ^ temp_ Value (?t) ^ diagnosis (temp, ?t) ^ has_value (?v,?h) ^ swrl:greaterThanOrEqual (?h, 125) −> has_disease(? x,fever)
2. Patient (? , x) ^ has_sensor (?x, B_Glucose) ^ B_Glucose _ Value(?t) ^ diagnosis (B_Glucose, ?t) ^ has_value (?v,?h) ^ swrl:lessThanOrEqual (?h, 0.7) −> has_disease(? x, Hypoglycemia)
3. Patient (? , x) ^ has_sensor (?x, B_Glucose) ^ B_Glucose _ Value(?t) ^ diagnosis (B_Glucose, ?t) ^ has_value (?v,?h) ^ swrl:greaterThan (?h, 0.7) ^ swrl:lessThan (?h, 2.5) −> has_disease(? x, NormalGlycemia)
4. Patient (? , x) ^ has_sensor (?x, B_ressure) ^ B_Pressure _ Value(?t) ^ diagnosis (B_Pressure, ?t) ^ has_value (?v,?h) ^ has_symptoms (?p,?systolic) swrl:greaterThanOrEqual (?h, 140) −> has_disease(? x, HighBloodPressure)
5. Patient (? , x) ^ has_sensor (?x, B_pressure) ^ B_Pressure _ Value(?t) ^ diagnosis (B_Pressure, ?t) ^ has_value (?v,?h) ^ has_symptoms (?p,?diastolic) swrl:greaterThanOrEqual (?h, 90) −> has_disease(? x, HighBloodPressure)

Rule set 2:

1. Patient (? , x) ^ has_sensor (?x, B_ressure) ^ B_Pressure _ Value(?t) ^ diagnosis (B_Pressure, ?t) ^ has_value (?v,?h) ^ has_symptoms (?p,?systolic) swrl:greaterThanOrEqual (?h, 180) −>has_disease(? x, Hypertensive)

2. Patient (? , x) ^ has_sensor (?x, B_ressure) ^ B_Pressure _ Value(?t) ^ diagnosis (B_Pressure, ?t) ^ has_value (?v,?h) ^ has_symptoms (?p,?diastolic) swrl:greaterThanOrEqual (?h, 120) −>has_disease(? x, Hypertensive)
3. Patient (? , x) ^ has_sensor (?x, Heart_Beat) ^ Heart_Beat _ Value(?t) ^ diagnosis (Heart_Beat, ?t) ^ has_value (?v,?h) ^ swrl:greaterThan (?h, 102) ^ swrl:lessThan(?h, 150) −> has_disease(? x, Fast_Heart_Beat)
4. Patient (? , x) ^ has_sensor (?x, Heart_Beat) ^ Heart_Beat _ Value(?t) ^ diagnosis (Heart_Beat, ?t) ^ has_value (?v,?h) ^ swrl:greaterThan (?h, 40) ^ swrl:lessThan(?h, 58) −> has_disease(? x, Slow_Heart_Beat)
5. Patient (? , x) ^ has_sensor (?x, B_Glucose) ^ B_Glucose _ Value(?t) ^ diagnosis (B_Glucose, ?t) ^ has_value (?v,?h) ^ swrl:greaterThan (?h, 2.5) −> has_disease(? x, Hyperglycemia)

Rule set 3:

1. Patient (? , x) ^ has_symptoms (? x, Hyperglycemia) ^ has_Symptoms (? x, HighBloodPressure) −>has_Diagnosis (? x, Diabetes)
2. Patient (? , x) ^ has_symptoms (?x, Hypoglycemia) ^ has_Symptoms (? x, HighBloodPressure) −> has_Diagnosis (? x, Diabetes)
3. Patient (? , x) ^ has_symptoms (? x, Fast_Heart_Beat) ^ has_BMI_value (?x, high) −> has_Diagnosis (? x, Diabetes)
4. Patient (? , x) ^ has_agevalue (? x, <35) ^ has_BMI_value (?x, high) ^ has_ symptoms (?x, Hypoglycemia)−> has_Diagnosis (? x, Diabetes)
5. Patient (? , x) ^ has_agevalue (? x, >45) ^ has_BMI_value (?x, Medium) ^ has_symptoms (?x, Hypoglycemia) ^ has_ symptoms (?x, HighBloodPressure)−> has_Diagnosis (? x, Diabetes)

The above discussed rules are samples. Similarly, fifty rules are constructed for all attributes which are considered in the ontology. The rules of set 2 detect the class label of the given data into two as: diabetic or non-diabetic. Each set contains several numbers of rules to handle the heterogeneity in the collected data. Finally, the prepared data are stored in an RDF format and then given to the classifier for classifying the results. The following algorithm 2 is used for generating rules.

Algorithm 2: Rules Generation

Input: Semantically enriched data

Output: Validated Data

Define R_s = new int [n] // **R_s means Rule Set**

R_s = {R_{s1} , R_{s2} , R_{s3} , R_{s4}, ….. R_{sn} } // **Rules set**

O = {Sub, Pre, Obj}

For each d_i in O // **each data(d_i) from ontology**

{

Check R_s against O // **Compare the data values with rule set**

for (int i=0; i<n ; i++)

if (n == true) // **stop if rule is true**

{

Rule n validated successfully;

}

else // **continue until all rule fails**

{

Rule execution error;

}

end else

end if

end for

end

C. Classification Model

The first and foremost step in the deep network model is to define the count of both the inputs (I_{in}) and hidden layer (H_i) along with its corresponding bias (b_i) and weights (w_i). The input layer is constructed with 14 number of nodes. In the first, the input layer is prepared where diabetes ontology data proceed to the 14 nodes as the input. Secondly, hidden layers are created with 200 neurons, wherein as the number of hidden layers increases, the accuracy of the results also increases. So, three hidden layers are developed and used in the proposed model. The semantically improvised data is fed to the input layer of the network and the defined input function in the network I_{in} triggers the merging of all inputs with its respective weights as shown in the following eq. 4.1.

Let us say d_1.....d_i as the inputs,

$$I_{in} = \sum_{1}^{n} d_i w_i + b_i \qquad \text{... Eq (4.1)}$$

The above eqn 4.1 is used for determining the inputs, where d_i and w_i denote the inputs and their corresponding weights respectively. Also, w gives the data pertaining to the inputs. A weighted sum is calculated for each neuron, which is the output value of each neuron in the adjacent network layer and it is then multiplied by their respective weight of the connection with that neuron. As I_{in} passes through the first hidden layer, the output of the first hidden layer is obtained as follows in the below eq. 4.2

$$I_h = ActiF\left(W_{iI_{in}} + b_i\right) \qquad \text{... Eq (4.2)}$$

where W_i is the weight matrix between the input and first hidden layer, b_i is the bias, and ActiF () refers to the activation function, which is designed to create a neural network model, wherein the usage of Nonlinear and multi-layer neural networks make sense. In the DNN model, many activation functions are available. Here, the ReLU is a basic activation function, it's selected to activate the output of the n^{th} hidden layer. The output of the n^{th} hidden layer can be obtained by using the eq. 4.3.

$$I_{h_n} = \mathrm{Re}\,LU\left(W_{h_n} I_{h_{n-1}} + b_{h_n}\right) \qquad \text{... Eq (4.3)}$$

The output value y can be expressed as a function of the input values and network weights.

It obtains inputs by multiplying w for each neuron and then generates an output that is relative to the input vector. The output Y is formed by applying the **softmax method** to obtain the predictive value of the network, i.e., given in the eq. 4.4,

$$Y = Soft\max(W_O I_h + b_O) \qquad \text{... Eq (4.4)}$$

First of all, data are trained and further, it will be tested with the help of the network prepared. The working steps of this algorithm of dealing with the IoT semantic-oriented data are given below.

Working Steps

Step-1: Elucidate a neural network consisting of an input layer as I_{in} that has input nodes, defined as 'n'.

Step-2: Elucidate and then initialize the count of hidden layers as I_H, which is required for training the data.

Step-3: Elucidate the bias value for each node and then its corresponding weight have to be chosen randomly in the beginning of the forward propagation stage.

Step-4: Elucidate the activation function for the hidden layer that is in the end.

Step-5: Apply the Rectified linear unit **ReLU ()** for acquiring the resultant of the output layer

Step-6: Train the network for the considered training dataset.

Step-7: Once the network finishes training, sweep the test data to the network that was trained for determining the models' classification rate.

Step-8: Alter the output result by using the method of softmax for acquiring the prediction value.

Step-9: Estimate the accuracy of the model with the help of various evaluation metrics.

Algorithm 3: Deep neural network model

Input: Ontology Data

Output: Classification of data

Define $I_{in} \leftarrow O$; // **I_{in} Input Layer, O Ontology data assigned as the input**

Define $I_H \leftarrow H_{1...n}$ // **I_H Hidden Layer, H is a number of hidden layer**

Define $W_i \in W$ //**weight matrix between the input layer and the first hidden layer, where i= (1,2, 3...,n)**

Define $b_i \in b$ // **Bias vector, where i= (1,2,3,...,n)**

Define I_o; // **I_o output Layer**

$I_{in} = \sum_{1}^{n} d_i w_i + b_i$; // **input layer includes all input with its weight and bias**

f_n = ReLU (); // **activate the function to designed model as meaningful**

I_H = fn ($W_i d_i$+ b_i) // **output of the i^{th} hidden layer is obtained**

Y = softmax ($W_{out} d_h$ + b_{out}); // **transformation of the output result by the softmax method to obtain the prediction value**

End

4.5.3 Phase 3: Result Evaluation

The results of the proposed model have been investigated in the results section. The most commonly used parameters, namely: precision, recall, accuracy, and F score are being computed to assess the accuracy of the developed RESOC_DNN model. The accuracy of the proposed model is compared with other traditional models like RNN, and ANN. The detailed description of the result section is described in the following section.

4.6 Results and Discussions

The results derived by the above-proposed model are described in this section. The correctly classified data are analyzed with the proposed model and labelled into two as: diabetic and non-diabetic based on the rule embedded ontology. Thus, rules embedded ontology is analyzed using the proposed algorithm.

Fig 4.4 represents the percentage of classified labels in the data. Positive class labels (i.e.) diabetic labels with the percentage of 88 and non-diabetic labels with the percentage of 12 are obtained.

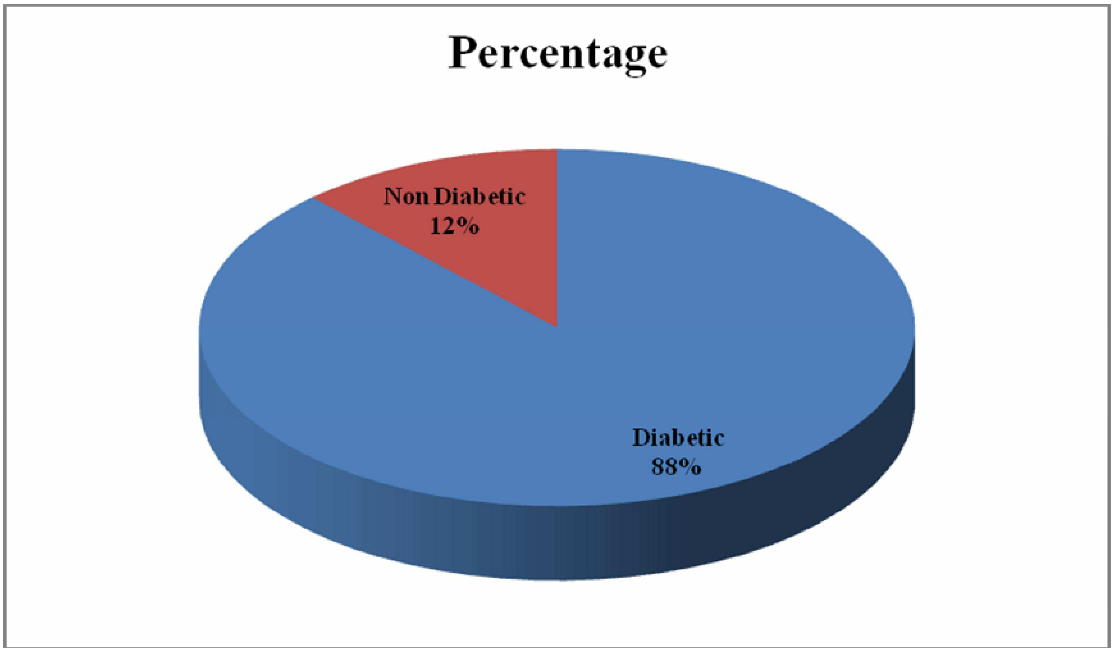


Fig. 4.4 Correctly Labelled classes in percentage

Fig. 4.5 represents the comparative analysis of the proposed model with existing models. The X-axis represents the features and Y-axis represents the count of data. The correctly classified data are labelled into two as: diabetic and non-diabetic with counts, which are represented in the following figure 4.5.

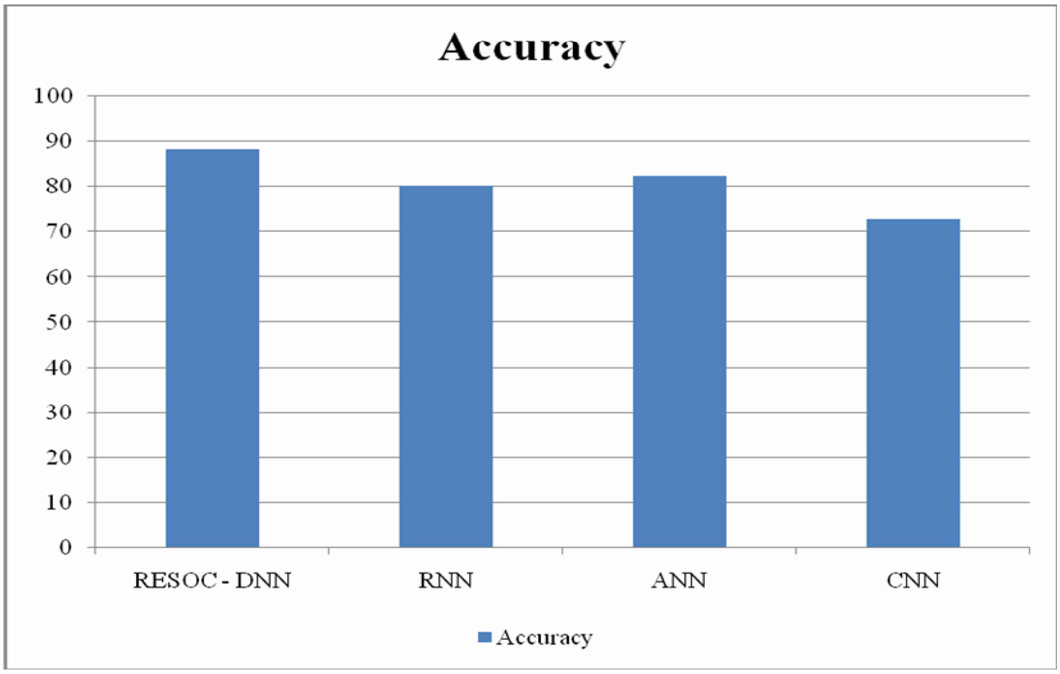


Fig. 4.5 Comparative Results of Classification Accuracy

Fig. 4.6 represents the precision, recall, F1-score, and accuracy of the collected data. The X-axis represents the common measures and Y-axis represents the percentage of common measures. Precision and recall produced a moderate result. This showed the efficiency of the proposed approach.

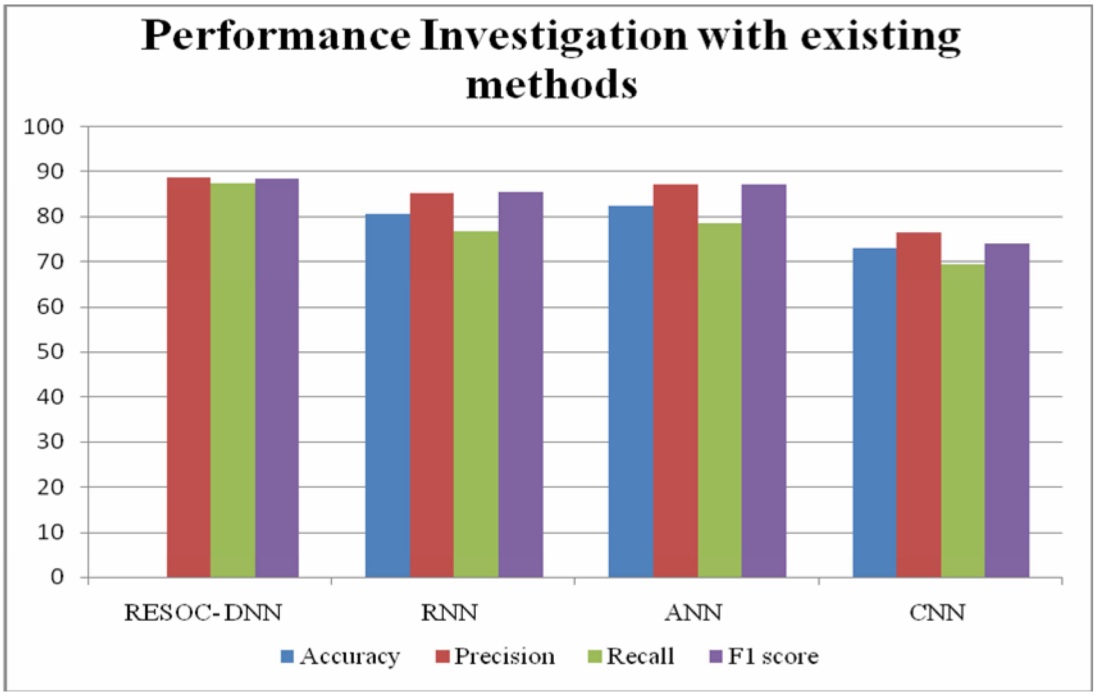


Fig. 4.6 Measurements of Precision, Recall, F-score and Accuracy

The proposed RESOC_DNN model results in 88.33% of accuracy, which is higher than the existing RNN, ANN and CNN models that were producing 80.35%, 82.36% and 72.88% of accuracy respectively. The proposed RESOC_DNN model is better by exhibiting 6% more accuracy than the existing models.

4.7 Chapter Summary

After reviewing the various literatures, the research which had been done on the analysis of IoT-based data for specific disease prognosis but still perfect model is needed for handling IoT data. The systematic diagram discussed was able to give a larger picture of the processing of IoT-based clinical data analysis. This chapter provided a newer ontology-based model for analyzing IoT-based clinical data. A novel RESOC_DNN along with ontology and various rules had been proposed to improvise the accuracy of the proposed model than the existing works. Adding rule embedded ontology into the model had resulted in an enhanced accuracy rate of 6% more than the RNN and ANN models. The proposed model predicted the diabetes disease accurately. Moreover, finding the risk level of diabetes could assist any patients in protecting themselves from a critical stage, which also reduces the workload of modern age doctors. But RESOC_DNN model could not be suitable for predicting the risk of diabetes. So, further enhancement of the model will be needed to predict the risk level of diabetes as well.

Chapter – 5

Improved-DNN with Binary Multi-Objective Optimization (IMO_DNN) Model for Predicting the Risk of Diabetes Patients

CHAPTER - 5

IMPROVED-DNN WITH BINARY MULTI-OBJECTIVE OPTIMIZATION (IMO_DNN) MODEL FOR PREDICTING THE RISK OF DIABETES PATIENTS

5.1 Introduction

In the investigation of IoT data, the prediction process is one among the most important processes under taken in it. In General, learning-oriented algorithms are deployed for synthesizing a prediction model. Likewise, this chapter devises a prediction model for the risk prediction of diabetes diseases on the basis of the synthesized ontology data that was detailed in the earlier chapter. The devised diseases prediction model deploys the deep neural network along with increased hybridization functions in order to improve the accuracy of it. In any IoT-based environment, predictive investigation is a crucial methodology for predicting any consequential events taking place in a domain-oriented data. Also, the chapter will be discussing the earlier researches that have considered the prediction process by deploying numerous of learning algorithms along with the proposed prediction model.

Further organization of this chapter is given as: In the background section, brief reviews on deep neural networks, Optimization and softmax layer will be discussed. The existing works on IoT data prediction using various algorithms is discussed in the related work section. The motivation and scope of the proposed work are elucidated in detail. The proposed prediction approach is illustrated in the “IMO−DNN Model” section. In the results and discussion section, the acquired results are analyzed and discussed. Finally, the overall summary of this chapter is described at the end of this chapter.

5.2 Background Study

The concept of deep learning was derived from the study of Artificial Neural Networks (ANNs) [Abi, 18]. ANNs is an active research area over the past few decades [Ade, 14], [Tka, 16], [Jia, 17b], [Wal, 18], [Gue, 20] and [Pan, 21]. Creating a quality Neural Network (NN) had to be ensured, since that was essential for the use and repair of neurons to generate truly valuable functions by adjusting the weights. However, depending on the issues in the NN, its training process can be time consuming with simple computational levels. Back Propagation (BP) was an efficient slope descent algorithm that had played a key role in NNs since the 1980s. BP prepares the ANNs with a supervised learning approach. Although training accuracy was high, there may be chances of getting regression performance in it when applying the experimental data. The back propagation was based on local gradient information consisted with a random starting point that often got stuck in the local optima.

Hinton [Hin, 06] proposed a new training method (layer-wise-greedy learning) in 2006, which marked the birth of deep learning techniques. The basic idea of layer-wise-greed learning was unsupervised learning, which would be done before the next layer-by-layer network pre-training. The data dimension was reduced and there was a small representation, which was obtained by extracting features from the inputs. Then, if the features were exported to the next layer, all models would be labelled and network would become well formatted with named data. The reasons for the popularity of deep learning were two fold. On the one hand, the development of large data analysis techniques indicated that the problem may be somewhat matched to the training data getting resolved and also the pre-training process could provide a random start before supervised learning data for the network. With the rapid development of computational techniques, a powerful framework was provided by ANNs deep architecture for supervised learning. In general, the deep learning algorithm consisted of multi-layered architecture with a non-linear information processing unit for each layer. Deep neural networks (DNNs), which use deeper structures in NNs, represented more complex functions as the number of layers and units in a layer started to increase. Given the adequately named training databases and appropriate models, deep learning approaches could help in establishing the mapping functions for operational convenience, which could offer the capability of optimization functions and softmax layer to predict the efficiency of the perfect class value[Sch, 15], [Wan, 16], [Was, 19] and [Yan, 21].

5.3 Related Works

Tawfik et al., [Taw, 21] designed a DNN based medical decision system to predict diabetes disease. This system used grid search hyper parameter optimization for DNN model performance evaluation. The proposed DNN system outperformed the five prediction models, namely Random Forest, Decision Tree, Extreme Gradient Boosting (XGBoost), Support Vector Machine, and Logistic Regression. The response time was increased and this system was not suitable for high-dimensional data.

Saiyed et al., [Sai, 21] extracted and categorized the histopathological images by using K-means Consensus Clustering-based CNN architecture. This CNN architecture used a deep neural network for model selection and recovery functions. This architecture suffered due to object position identification in the image, missing values, and outliers

Lu Men et al., [Men, 21] proposed an LSTM based deep learning approach for a multi-disease prediction system. LSTM approach was classified into two mechanisms, namely time aware and attention-based to handle multi-label classification. The time aware mechanism was used to perform multilevel classification and an attention-based mechanism was used to resolve the problems in the multi-label classification. Overfitting problem occurred due to training data with a minimum number of variables.

Rani et al., [Ran, 21] improved the performance of deep learning technique by adding bi-directionally Convolutional LSTM (BiConvLSTM). BiConvLSTM was used to connect all fully connected layers. The proposed technique was evaluated against three data sets namely DRIVE, STARE and CHASED DB1. Random weight was used to connect the fully connected layers which led to an increase in the system complexity (memory).

Shorif et al., [Sho, 21] proposed Sequential minimal optimization (SMO) and Artificial neural network (ANN) based hybrid method to forecast heart and diabetes disease. Multilayer perceptron (MLP) in SMO was to correlate the features. MLP in ANN was used to connect the input and output layers. SMO increased the time complexity which affected the ANN network.

5.4 Objective

From the above literature review, most of the work had been carried out by using the Deep learning algorithms and confusion matrix to find the accuracy of data. For the unsupervised data, K Means based model produced better results. Optimization of the prediction model and loss function minimization helps to improve the accuracy of the model. This is motivated the author towards the current work of building a new deep neural network model based on classification of groups using the Softmax for predicting the target class to enhance the accuracy.

The objectives of the proposed approach are:

- To enhance the accuracy using a deep neural network model for the collected dataset.
- To classify the groups of unsupervised data based on the unsupervised learning algorithm for the prepared dataset.
- To predict the target class with better accuracy.

5.5 IMO_DNN Model

The main objective of the proposed Improved-DNN with binary Multi-objective Optimization (IMO_DNN) model is to predict the risk of diabetes mellitus using prepared data and then improve the accuracy of the prediction using different layers with various functions like loss and optimization as a novelty of this work. With binary Multi-objective Optimization, the two objectives like reduction of the ratio of NSR- Noise-to-signal is done along with the optimization of parameters like accuracy, precision, F1 Score, and Recall. The dependability of using certain configuration hardware could be eliminated by using the multi-objective optimizations.

In the previous chapter, diabetes mellitus is forecasted where the class labels are named as “diabetic” and “non-Diabetic”. The prepared ontology data is given to this proposed model, but here, the risk of diabetes diseased person is predicted. It will help the patient to prevent them from a serious stage. Initially, the labels are unknown and the proposed model predicted and labelled the data as: low risk, high risk or medium risk as given in the following table.

Table 5.1 Risk assessment of Diabetes Mellitus

Low Risk	Medium Risk	High Risk
Age < 35	Age 35 to 40	Age > 40
Blood pressure < 90 mmHg	Blood pressure 90 to 120 mmHg	Blood pressure =>140 mmHg
Blood Glucose < 100 mg/dl	Blood Glucose 100 to 125 mg/dl	Blood Glucose => 126 mg/dl
Pulse rate 80-120 bpm	Pulse rate 120-130 bpm	Pulse rate < 70 && > 130 bpm
BMI < 18.5 kg/m^2	BMI 23 to 24.9 kg/m^2	BMI > 25 kg/m^2
Daily calorie < 1000	Daily calorie 1200 to 2200	Daily calorie > 2200
Physical Exercise = Twice a day	Physical Exercise = Once a day	Physical Exercise = No

5.5.1 Loss Function

Introduce the loss function given by the relation,

$$v = w'd + b \qquad \text{... Eq (5.1)}$$

Here, w' represents Weight Matrix; x represents a varying sample input in vector format; and b represents Bias vector. Since, both right hand side elements are in vector format, the resultant would also be a vector, denoted by v_i after defining $w_i^{'}$ and b_i as ith elements of w' and b respectively.

$$v_i = w_i^{'}d + b_1 \qquad \text{... Eq (5.2)}$$

Since the bias is negligible, it could be neglected and be simply rewritten as:

$$v_{gt} = w_{gt}^{'}d \qquad \text{... Eq (5.3)}$$

Here, gt will be the true class label of d and v_{gt} will be the product yield by multiplying both w_{gt} and d. To facilitate training process, we define regularizing function as given below to favor the aligning of the two vectors under discussion.

$$R_f = \left| w_{gt} - \vartheta \frac{d}{d'd} \right|_2^2 \qquad \text{... Eq (5.4)}$$

Here, ϑ represents the unit vectors' scalar co-efficient in the direction of d and the squaring of the regularizing function is found to aid in preventing overfitting issue. Now, merge them with the above function with two loss functions, namely marginal loss and Mean Square loss, which is given in eq. 5.5

$$Loss = \left(v_{gt} - 1\right)^2 + \sum_{p \neq q} \left(v_p - 0\right)^2 + \sum_p M\left(0, 1 - v_{gt} + v_p\right) + R_{f*S_p} \qquad \text{... Eq (5.5)}$$

Here, R_f is a scalar parameter, which could be founded on the validation set. Thus, if $|v_{gt}| = |w_{gt}^{'} d|$ tends to 1, ϑ could be fixed at 1 so as to avoid mis-classification and thereby the noise may not appear to be input while processing that data. There will be no marginal loss for mistakenly predicted samples, but for rightfully predicted samples, it has both R_f and marginal loss. Thus, the correct output of the methodology could be always ensured.

5.5.2 Optimization

After the application of loss function, the output from this function is given to a multiple-objective based Bayesian Optimization to optimize the classification process taken up to correctly predict the diabetes diseases risk level by ensuring objectives like Higher values of accuracy, precision, F1Score, and Recall.

In this multiple-objective based Bayesian Optimization, every objective will be approximated with the deployment of model of Gaussian method. For any specific objective function, f(d) is defined as a surrogate Gaussian method; $f_p = f(x_p)$ is defined as an objective function for considered network x_p; d_p as actual measured value of the function, which could be defined otherwise by $f_p = f(d_p)$; and $d_{1:n}$ as estimated network structure. The optimization equation is given by,

$$f \sim N(\xi, \phi), d \mid f, \varphi^2 \sim N(f, \varphi^2) \qquad \text{... Eq (5.6)}$$

5.5.3 Working Methodology

The structure of the proposed model will be same as the conventional deep learning technique that includes the input layer, hidden layer, and output layer. The input layer is constructed with 14 number of nodes. In the first, the input layer is prepared where diabetes ontology data proceed to the 14 nodes as the input. Secondly, hidden layers are created with 200 neurons, wherein as the number of hidden layers increases, the accuracy of the results also increases. So, three hidden layers are developed and used in the proposed model. These hidden layers are deeply trained and uniquely operate the data. In the diabetes ontology data, the risk of diabetes mellitus is unknown.

For each iteration, measuring similarity could empower the model to process and identify the similarities, which are done in the output layer. Output layer performs two important works, namely: matching and finding the class labels. The softmax activation function predicts the accurate results by determining and comparing each class probabilities and threshold values based on deep trained layers. There are two important processes training and testing. The following sections are described the training procedures of data.

5.6 Training Process of the proposed model

The proposed network model is designed with five different layers that are one input layer, three hidden layers, and one output layer. The subsequent sections explain the construction and process of these layers.

5.6.1 Input Layer

This is an initial layer in the network which is prepared with 'n' number of neurons for processing input. Input vectors are gathered from the diabetes ontology and then given to each neuron. In terms of ontology, the input vector can be called knowledge. The form of each knowledge like = {S, O, P}, where S − Subject, O –Object, and P − predicate. Input vectors from the knowledge base (i.e., ontology) are given to the input layer.

5.6.2 Hidden Layers

The input layer passes the K samples to the hidden layer. In the proposed model, three hidden layers are constructed. The first two hidden layers for selecting features and these are created to training the data. The third layer for data grouping and which is configured by the clustering technique where the similarity matching method is used. The quintessence from the output layer of the second hidden layer is given as the inputs to the third hidden layer.

5.6.3 Output Layer

The output layer will be the final layer in the used network, where in the softmax function will be deployed for acquiring more accurate outcomes.

A. Softmax function

The Softmax function is a linear classifier that makes probability based on logistic regression. It is also known as polynomial Logistics regression. Softmax is an improved version of logistic regression, which splits up the two classes as logistic regression. The Softmax can be used to classify two or more classes as the contrast for logistic regression. It is often used as an activation function to calculate the effect of a neural network. Softmax is affiliated with the output layer for predicting the features from the previous layer. The softmax function then evaluates the group of specific inputs entered in the output layer.

In the issue of multiclass, Softmax will be evaluating the probability of every data class getting classified. Therefore, the classes likelihood sum will be equal to unity (i.e.) 1. Further, SF executes the exponential and normalization processes for identifying the class likelihood. At final, the output will be predicted by yielding utmost accuracy with the deployment of the softmax function in the networks' final layer. The softmax function is defined as $\delta(d)_k$, in the output layer

$$\delta(d)_k = \frac{e^{d_k}}{\sum_{i=1}^{i} e^{d_n}} \qquad \text{... Eq (5.7)}$$

Here, d represents the inputs' vector onto the output layer and k indicates the count of output units, $k = 1, 2, ..., n \cdot e^{d_k}$ and e^{d_n} are the exponential functions of input and output vector respectively.

The working steps of the algorithm is given below,

Working Steps

Step-1: Elucidate a neural network consisting of an input layer as I_{in} that has input nodes, defined as 'n'.

Step-2: Elucidate and then initialize the count of hidden layers as I_H, which is required for training the data.

Step-3: Elucidate the bias value for each node and then its corresponding weight have to be chosen randomly in the beginning of the forward propagation stage.

Step-4: Apply the K means algorithm for grouping the data which get from the previous layer.

Step-5: Elucidate the activation function for the hidden layer that is in the end.

Step-6: Train the network for the considered training dataset.

Step-7: Once the network finishes training, sweep the test data to the network that was trained for determining the models' classification rate.

Step-8: Alter the output result by using the method of softmax for acquiring the prediction value.

Step-9: Then, make use of the loss function to avoid unnecessary noises in the data.

Step-10: Finally, integrate the output from the loss function to binary multi-objective optimization function, wherein the Higher values of accuracy, precision, F1 Score, and Recall are ensured to achieve parameter optimization in our work

Step-11: Estimate the accuracy of the model with the help of various evaluation metrics.

Algorithm: IMO_DNN model

Input: Classified Ontology Data

Output: Risk Prediction

Define $I_{in} \leftarrow O$; // **I_{in} Input Layer, disease classified Ontology data (O) assigned as the input**

Define $I_H \leftarrow H_{1...n}$ // **I_H Hidden Layer, H is a number of hidden layer**

Define $W_i \in W$ //**weight matrix between the input layer and the first hidden layer, where i = (1, 2, 3..., n)**

Define $b_i \in b$ // **Bias vector, where i= (1, 2, 3,..., n)**

Define I_o; // **I_o output Layer**

$I_{in} = \sum_{1}^{n} d_i w_i + b_i$; // **input layer includes all inputs with its weight and bias**

Apply K means algorithm; // **identify the class label by grouping nearest data items**

I_H = fn ($W_i d_i$+ b_i) // output **of the ith hidden layer is obtained**

Y = softmax ($W_{out} d_h$ + b_{out}); // **transformation of the output result by the softmax method to obtain the prediction value**

Apply Loss function;

Apply binary multi-objective optimization function;

End

5.7 Results and Discussion

The IMO_DNN hybrid model was proposed for the prediction of the risks involved in the diabetes disease. The proposed model is being evaluated and compared based on its performance. Here, Accuracy, Recall, Precision, and F1-score are estimated by using the below shown equations for evaluating the proposed model performance and then to make a comparative study of the results taken from other models like RNN, CNN, and ANN. The Accuracy, Recall, Precision and F1-score are being estimated by the deployment of the confusion matrix as given in chapter 3 (Eq. 3.1, 3.2, 3.3 and 3.4).

The dataset used for this comparative study was divided into two sets, namely: training and testing. 70% of the data in the set was designated to the model to train and the remaining 30% of the data in the set was utilized for model testing. The network model containing deep layers had been devised to predict the risk of diabetes disease. The results acquired from the proposed methodology is being compared with the following earlier models as tabulated in the below table 5.2.

Table 5.2. Comparison result of proposed model with traditional models

Evaluation Metrics	Accuracy	Precision	Recall	F1 score
IMO_DNN	92.27	93.67	90.93	92.28
RNN	80.35	85.12	76.63	85.21
ANN	82.36	87.01	78.52	87.11
CNN	72.88	76.51	69.23	73.74

In the table 5.2, few of the existing models like RNN, CNN, and ANN were chosen and used for the comparative study by taking the performance measures like Accuracy, Precision, F1 score, and Recall.

The proposed model predicted the diabetes risk with 92.27% of accuracy, while the other models like RNN, CNN, and ANN gave rise to lesser accuracies of 80.35, 72.88, and 82.36 respectively. The same has been depicted in the below fig. 5.1.

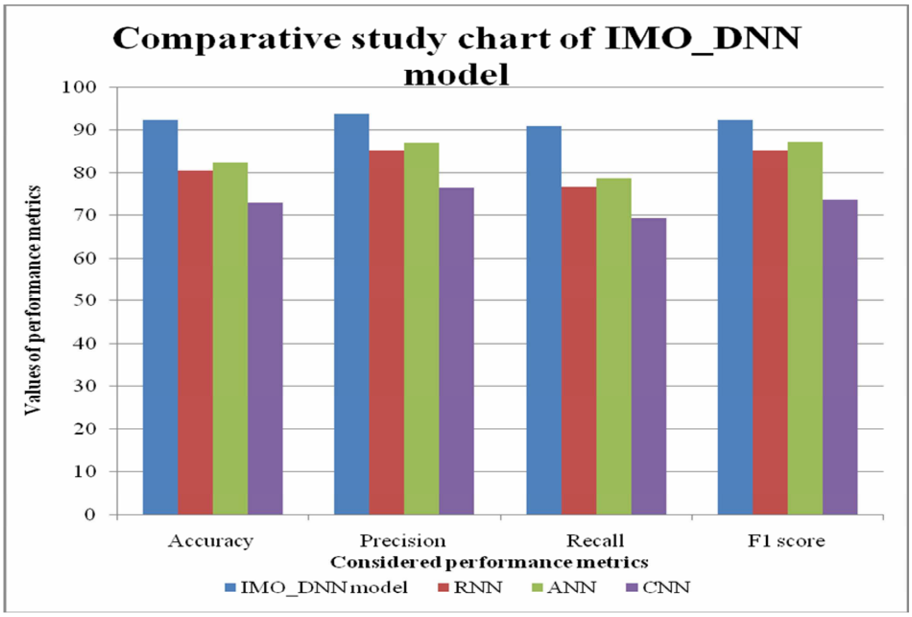


Fig. 5.1 Comparison of the proposed model prediction accuracy with other models

Furthermore, the proposed IMO_DNN model not only yielded better accuracy, but it also yielded better precision, F1 score, and recall values than the other models taken for the comparative study. While taking precision, the proposed IMO_DNN model yielded a higher precision value of 93.67% than the models like RNN, CNN, and ANN with precision values of 85.12, 76.51, and 87.01 respectively.

When comparing the F1 score, the current IMO_DNN model was found to produce increased F1 score of 92.28 than the F1 scores of 85.21 by RNN; 73.74 by CNN; and 87.11 by ANN. Finally, recall value produced by the IMO_DNN model was also found to be higher at 90.93% than the existing models like RNN, CNN, and ANN, which yielded lesser recall values of 76.63, 69.23, and 78.52.

From the comprehensive experimental results, the proposed model was able to outperform all other models that were taken for comparison. Moreover, the proposed model also followed a similar structure to a conventional network, but with the network training process and working methodology being drastically different.

5.8 Chapter Summary

In this current work, the Deep Neural Network-based model was devised to predict the risk of diabetes disease. The proposed model was built with the deep network that trained features by using three hidden layers. The IMO_DNN model constructed with Optimization and loss function for improving the accuracy of the model. The prepared ontology knowledge of collected data had been given as an input. Finally, the output layer predicted the exact result like low risk or high risk or medium risk. The experimental results were compared with other few existing models from the past researches in the similar arena.

It is evident that the results of the model outperformed better by yielding values of accuracy of 92.27%, precision of 93.67%, F1 score of 92.28 and recall of 90.93%. Hence, the proposed model was able to predict the risk of diabetes disease in an effective way. However, using the separate phase for selecting the features would have furthermore reduced the computation time and speed by eliminating the need for processing of unnecessary data to classify and label the risk. Thus, we introduce the XGBoost Gradient for working on the computation time and speed.

Chapter – 6

An Enhanced IMO_DNN (EIMOXB_DNN) Model with Booster to Improve the Accuracy of Risk Prediction

CHAPTER - 6

AN ENHANCED IMO_DNN (EIMOXB_DNN) MODEL WITH BOOSTER TO IMPROVE THE ACCURACY OF RISK PREDICTION

6.1 Introduction

One of the main problems and challenging tasks in the process of predictive analysis was to improve the accuracy of the model. To redress this problem, a new model, called the EIMOXB_DNN model is being proposed by combining the previously proposed work of the IMO_DNN with XGBoost Gradient approach. The proposed model aims to improve the accuracy that is discussed in this chapter. Diabetes ontology data are analyzed using the proposed IMO_DNN model and then the class labels are predicted accordingly.

Further organization of this chapter is given as: In the background, the boosting techniques are explained with reference to the many works in the literature. In the Related Works section, a review of some related works which related to diabetes data analysis using the state-of-the-art machine learning methods is briefly discussed. The main purpose of this proposed work is given in Section 6.4 by defining the objectives, which are aimed to be achieved, as found in 6.5. The working methodology of the currently proposed EIMOXB_DNN is being discussed in 6.6 and training phase of the work is presented in 6.7. Finally, the comprehensive result comparison is being made with earlier phase work of IMO_DNN and other existing methods in 6.8.

6.2 Background Study

Classification and prediction applications making use of learning models have started to gain much attention. Thus, diversified deployment of such classification and prediction approaches might sometimes end up in yielding few limitations. These limitations were even realized in the earlier reported techniques of image processing [Guo, 18] [kam, 19], text and document classification [Mir, 18], intrusion detection [Ali, 18b], clinical diagnosis [Sah, 20] [Aro, 19b], and format recognition [Eun, 20], all of which were found to deploy class labels that's of either binary or multiple classes.

Ever since the emergence of large database comprising of many features [Lia, 18], [Lal, 20]. This quantum of large data that got acquired had started to yield us with significant knowledge and understanding, provided that the data could be properly analyzed. As we gain insights in handling these large quanta of data, it's becoming a necessity to upgrade from machine learning approaches to deep learning. This could lead to significant advances in areas such as computer vision [You, 18] and classification problems [The, 17]. Furthermore, few traditional trivial models as in [Goh, 20] and few deep learning methods as in [Zan, 18], [Par, 19] found to deploy only one model for training. The benefits yielded out of such single model may not be adequate enough to achieve the required accuracy for complex problems. If plenty of models could be integrated and used up to its full potential, then benefits yielded out of it would surpass the limitations, which aids towards the sorting out of complex problems.

6.2.1 XGBoost algorithm

In recent years, the XGBoost algorithm [Che, 16] was widely liked by machine learning enthusiasts and practitioners due to its fast calculation speed and good model performance. It was a process of gradient enhancement with the advanced implementation of the algorithm that has been used successfully for some studies [Xia, 17], [Zie, 16].

It could help in regulating the overfitting-underfitting issues. Furthermore, it aids to increase the accuracy and yield faster computation speed, which was possible because of parallel usability of CPUs of all the hardware. Using the AUC of the ROC chart, a set of hyper parameter values could also be provided by the user.

In general, the modification of parameters in machine learning was often made [Dai, 11]. For instance, daunting task of controlling the knowledge rate in samples was one such parameter modification. Out of the many deployed parameter modifiers, eXtreme Gradient Boosting (XGBoost) [Che, 16] was found to yield higher accuracy and sophisticated performance. Having won many machine learning competitions, XGBoost was efficiently able to sort out over fitting by its boosting. Also, several hyper parameters that were problematic could be sorted out by upgrading a package hyper parameters with the deployment of XGBoost models. The appropriate alteration of such hyper parameters requires proper awareness in factors like training, general validation, and error detection whenever implanting an algorithm.

Many optimizations as in [Ber, 12] [Man, 15], and [Sno, 12], the tuning of hyper parameters in manual mode was very unsuccessful and unhopeful if number of hyper parameter values were higher. For instance, grid and random search were the methods that only trust the knowledge learned via the model with prior optimizations.

6.3 Related works

The diabetes disease prediction aims to prevent the diseased person from entering a critical stage by earlier or precise detection of the diabetes diseases. Some of the works reported in the literature encounter these aims, which will be discussed below.

A machine learning- oriented predicting methodology was suggested by [Kop, 20] with the deployment of multiple concepts like Glmnet, LightGBM, and XGBoost. They had considered the cases of Type 2 categorical Diabetic diseased person, wherein their plasma glucose levels were recorded after certain period of prescribed fasting period. They conducted around hundred iterative operations to validate the performance of each model deployed by them and found that the most effective model was LightGBM, surpassing the two other models of Glmnet and XGBoost. However, interpreting such complex models were difficult, which could in turn affect decision made out of results obtained from such models.

Another machine learning-oriented diabetes diseases classifying and predicting Logistic regression (LR) was done by [Man, 20a]. They made use of four diverse classifiers, namely: Adaboost, Naïve Bayes, Random Forest, and Decision tree by taking the 2009-2012 dataset published by National Health and Nutrition Examination Survey. This dataset consisted of a total of 6561 members, out of which 5904 were in control of the glucose levels and 657 were not in control of the glucose levels (i.e.) Diabetes diseased patients. They also deployed three types of partition protocols like K_2, K_5, and K_{10}. They didn't specifically consider the diabetic patients, which is the major drawback of it.

A comprehensive experimental investigation of deep and machine learning methodologies was done by [Sah, 20] with the deployment of Convolution Neural Network (CNN) as a feature extracting tool. They took the common machine learning methods like Logistic regression, K-Nearest Neighbors, Multi-Layer Perceptron (MLP) classifier, Random Forest, Decision tree, and Support Vector Machine (SVM) to compare with their feature extracting CNN method. They concluded that their CNN method was better than all other taken models in terms of accuracy. But their work was limited only to females, who were evaluated whether they are diabetic or non-diabetic.

A clinically aware machine learning system was introduced by [Jia, 2020] in order to forecast the mortality rate of the critically suffering diabetes patients. Many terminologies like Domain knowledge, Concept Unique Identifiers (CUIs), Unified Medical Language System (UMLS) entity, and CNN-oriented word embedding, and various rules and features were integrated successfully to classify the mortality rate. Their method was found analyzing the Intensive care data based on the knowledge of clinical experts and resources of UMLS to yield a promising accuracy of around 97 percentage. However, the system was too dependent on the clinical expert opinions to build the knowledge domain.

A general purpose critical disease aware system with modified SVM was proposed by [Har, 21] by taking critical conditions like Diabetes, Chronic Disease ailment, and Cardiac ailments. The data pertaining to the critical conditions were reduced and deployed to predict the disease, which in turn aided the doctors to provide personalized treatments to different patients. They also made a comprehensive investigation of their critical disease aware system with conventional models like SVM-Polynomial, Decision tree, SVM-Linear, and Random Forest on R-studio. Out of the three critical conditions, a promising accuracy of 98.7 percentage was obtained for predicting the Diabetes disease.

Features aware diabetic disease investigation was made by [Ahm, 21] by using the machine learning method in the health care. The features considered for their investigation were FPG and HbA1c. Not only these features were investigated, but also their impact on the diabetes disease was also made by them. They made use of five diverse classifiers and processes like feature permutation, feature elimination, and hierarchical clustering to successfully predict and realize the impact of the features. This model was geographically limited to the population of Saudi Arabia and might become void to different geographical areas.

[Isl, 20] proposed yet another machine learning-oriented diabetes diseases identifying and classifying methodology with special consideration given to the country of Bangladesh. They have taken the data from Bangladesh-based health survey of 2011, which consisted of 1,569 members, out of which 127 were suffering from diabetic condition and the rest were non-diabetic. Six classifiers such as SVM, logistic regression, bagged classification, linear discriminant analysis, linear discriminant analysis, and k-nearest neighbor were used to know the significant factors contributing to the diabetic condition. However, some of the few known limitations of the work were: genetical disease, circumference of waist, geographical dependence, and obsolete 2011 data, thus failing to be up-to-date.

Another instant of work concentrating on the critically death causing diseases like Cardiac ailment, cancer ailment, and diabetes diseases were done by [San, 20] to develop a prediction system using the temporal CNN and incremental feature selection. Better results in terms of accuracy, time, and false alarm rate were achieved with the deployment of the dataset taken from UCI. This work is found to be limited as it the data gathering was limited to an institution.

An innovative way of machine learning feature selection methodology was devised by [Le, 20] with the deployment of Adaptive Particle Swam Optimization (APSO) and Grey Wolf Optimization (GWO). It was helpful in reducing the count of the input features and earlier detection of diabetes diseases. They compared the performance of their proposed feature selection system other models like SVM, KNN, Random Forest, Naive Bayes, Decision tree, and Logical Regression. They were able to infer that only a few features would be needed for detecting the diabetes diseases. However, optimization of the features selected could have contributed towards an improved performance.

Another combination of set of diverse machine learning classifiers like KNN, Random, forest, Multilayer Perceptron (MLP), XGBoost, Naive Bayes, and AdaBoost were used by [Has, 20] towards the forecasting of the diabetes diseases. Weighted ensemble of these models was used to improvise the results of diabetic diseases forecasting by tackling issues like outlier rejection, data standardization, missing values, and K-fold cross-validation, etc. But its versatility and generality might be a question in healthcare without proper implementation.

6.4 Objective

From the above literary review, it is revealed that most of the works were done separately using machine learning and deep learning technique. Very little research was done on the IoT environment by using knowledge base for disease prediction. The optimized model for training will improve the accuracy of the prediction model. This prompted the creation of a new enhanced model based on combining XGBoost and Deep neural network methods to improve accuracy. The objectives of the proposed model are as follows:

- To provide an innovative hybrid-based model using Xgboost and Deep Neural Network concept.

- To predict the risk of diabetes disease using knowledge base in IoT environment.
- To improve the accuracy of the prediction model than the current method.

6.5 EIMOXB_DNN Model

The enhanced IMO_DNN (EIMOXB_DNN) model is being proposed for predicting the risk lying in the diabetes disease. This enhanced IMO_DNN model combines the performance of earlier proposed IMO_DNN and optimized eXtreme Gradient Boosting by introducing the novelty of integrating the feature selection with Adam optimizer.

The enhanced model consists of a web with multiple stacked layers and is the last layer of the XGBoost layer. This is different from conventional DNN because it does not have a fully integrated (FC) layer.

This initiates the simplicity and diminishes the number of calculation parameters because there is no need to bring back weights from FC layers to readjust the weight on the previous layers. Enhanced IMO_DNN model predicts the target labels with greater accuracy than any two individual models could do, for instance, individual models like Decision Tree [Pat, 18], MLPand Support Vector Classification (SVC) [Pra, 18].

6.6 Working Methodology

Various phases that used in the enhanced IMO_DNN are detailed below.

6.6.1 Feature Selection

Firstly, feature selection using the technique of XGBoost, which was scalable learning method deployed for tree boosting [Che, 16]. A feature significance score will be calculated for each contained decision trees in it for selecting best possible features in the data. Thus, the following equation for generating the features with the help of set of decision trees:

$$fv_k = \sum_{dt}^{dt} dtp_k \qquad \text{... Eq (6.1)}$$

Here, dt represents count of decision trees, fv_k represents the feature vector at the point of k^{th} data, and p_k indicates the prediction made from each decision tree. For facilitating the training process, it becomes indispensable to use loss function. The loss function that uses is given in the following equation:

$$Loss = -\frac{1}{N}\sum_{k}^{N}((fv_k \log p_i) + (1 - fv_k)) \qquad \text{... Eq (6.2)}$$

The role of Regularization could be significant in case of XGBoost system, which could be given by,

$$R_f = r^l + \frac{1}{2}\lambda\sum_{i=1}^{l} S_i \qquad \text{... Eq (6.3)}$$

Here, l represents count of leaves and S_i represents the score obtained on the i^{th} leaf. Further, the objective of the system could be depicted as:

$$O = Loss + R_f \qquad \text{... Eq (6.4)}$$

Here, $Loss$ function indicates training loss that how effective the prediction is made and R_f is nothing but the regularization function that tackles the issues of overfitting and maintains the complexity of the model.

Generally utilize the variance and mean in the XGBoost for the purpose of optimizing the defined objective function as follows:

$$O(j) = \sum_{k=1}^{n}\left[a_k dt_k^p(fv_k) + \frac{1}{2}b_k dt_k^p(fv_k)\right] + R_f(fv_k) \qquad \text{... Eq (6.5)}$$

The following steps are concerned when combining XGBoost,

Step 1: For n count of times, construct a tree till it attains the peak depth (every node will be indicating each feature). Then, determine the appropriate point of splitting, which is followed by the assigning of weight to the two newer leaves.

Step 2: End for- Trees in the model will get constructed. Then, define the feature set, v_k as $v_1, v_2, ..., v_n$. Here, k will be equal to values of 1, 2, …, n. Afterwards, the score of feature significance Fs_k is estimated on the basis of the frequency of each feature getting utilized for training the data across every trees. The score is given by,

$$Fs_k = \left\{ \frac{d}{d} = w_k v_k \right\} \quad \text{... Eq (6.6)}$$

Here, v_k represents the feature set itself and w_k indicates the weight of every feature present. This score will be useful when it comes to experimental validation, since XGBoost was earlier used an single classification model only [Li,20][Che, 18] and [Dha, 18].

6.6.2 Prediction

Then, the classification gets started with the defined DNN, a feed forward neural network. The input layer in the neural network will be feeding the ontology as an input to the model. In the network, as the count of input neurons is identical as the input features of the dataset, the input layer with n number of inputs is given by,

$$D = [d_1, d_2 - d_n] \quad \text{... Eq(6.7)}$$

Since the addition of more hidden layers are feasible in DNN, which define the input of the hidden layer that can map the input D from the input layer with a bias, v_k and random weights, w_j as

$$h_k = \sum_j w_j v_j + v_k \qquad \text{... Eq(6.8)}$$

Here, k represents the count of hidden units on DNN, which can vary in the range of 1, 2, 3… K and v_j represents the Feature vector in the data point of j.

Any hidden layer could be related with a function of non-linear activation. One such is ReLU, which removed the issues of exploding gradient and vanishing. Thus, the aim to apply ReLU and further, the results of hidden layer are related and given as,

$$H = f(h_k) \qquad \text{... Eq(6.9)}$$

where,
$$f(h_k) = \mathrm{Re}\,LU(h_k) \qquad \text{... Eq (6.10)}$$

As a result, the output layer generates the output of DNN as given in the below equation 6.11 as it processes the inputs from the predecessor hidden layer to the activation of the output layer. For intrusion identification and transforming inputs into probability classes $\delta(d)_k$, that uses softmax function in the output layer.

$$\delta(d)_k = \frac{e^{d_k}}{\sum_{i=1}^{i} e^{d_n}} \qquad \text{... Eq(6.11)}$$

Here, d represents the inputs' vector onto the output layer and k indicates the count of output units, $k = 1, 2, ..., n \cdot e^{d_k}$ and e^{d_n} are the exponential functions of input and output vector respectively. Then, the training of the neural network with the dataset takes place and finally, fitting and over fitting will be decided.

6.6.3 Optimization

The training of the network might get affected by the rate of learning, which is tackled by the usage of Adam optimizer. The deployment of Adam optimizer makes use of values of variance and means to maintain this rate of learning. Also, to reduce the chances of errors in any study, the selection of appropriate network structure and parameters should be made. Now, the explanation of the deployment of Adam optimization onto IMO_DNN as follows:

Declare: α Stepsize

Declare: $\beta^1 \in (0,1)$

Declare: $\beta^2 \in (0,1)$

Declare: f(x) // **objective function (detection accuracy maximisation)**

Declare: δ_v : parameter vector v

$m_v^1 = 0$ // **First movement vector**

$m_v^2 = 0$ // **Second movement vector**

$t_s = 0$

$t_s = t_s + 1$

While $f(x)$ do

$\varphi_{t_s} = \coprod \delta f_{ts}(x_{t_s} - 1)$ // **gradient of objective function at** t_s

$m_v^1 t_s = \beta^1 * m_v^1 t_{s-1} + (1-\beta^1) * \varphi_{t_s}$ // **Update first bias for first movement estimation**

$m_v^2 t_s = \beta^2 * m_v^2 t_{s-1} + (1-\beta^2) * \varphi_{t_s}$ // **Update first bias for first movement estimation**

$\overline{m_v^1 t_s} = \dfrac{m_v^1 t_s}{(1-\beta^1)}$ // **calculate the first bias in the estimation of the first movement.**

$\overline{m_v^2 t_s} = \dfrac{m_v^2 t_s}{(1-\beta^2)}$ // **calculate the second bias in the estimation of the second movement**

$x_{t_s} = \dfrac{\overline{m_v^1 t_s}}{\sqrt{m_v^2 t_s - \in}}$ // **update parameters**

End while

Return x_{t_s}

6.7 Training process of the proposed model

The proposed model contains two main parts, each part is made up of different layers that are the feature learning part and the prediction part that does the classification with loss function and Adam optimizer. The Adam optimizer ensures appropriate setting of the learning rate by avoiding the chances of encountering errors. The feature learning part has three layers, namely: the input layer and hidden layers with XGBoost layer. The prediction part contains the prediction layer and output layer. Each layer has different skills and responsibilities and these layers are the keys to the success of the model.

The Feature learning part learns key features from the training of the knowledge base. Different layers will be present in a similar way that the previously proposed model had in it.

6.7.1 Input layer

The input layer is the initial layer and is the sample input of the model. The training data K driven from the knowledge base will be given to the input layer. K consists of a set of rows as (x_i, y_i), where i is the key of the data, x_i is the feature matrix, and y_i is the target class. If the training model is in this required format, it will be sent directly to the hidden layers of the feature learning section, otherwise, it will be formatted in the data pre-processing layer. The input vectors are sent to the data preprocessing layer to get finely tuned data.

6.7.2 Hidden layers

The hidden layers are introduced after the input layer to do the non-linear conversions as and when needed. ReLU activation is being introduced before the data gets shifted to the output layer.

6.7.3 Optimized XGBoost layer

Both methodologies of XGBoost and gradient boosting were ensemble tree processes that deploy the structure of gradient descent for boosting the weaker learners. Although, XGBoost build up the basics of Gradient Boosting architecture by optimizing the system and algorithm, its development was slower. XGboost is a package owned by the Distributed Machine Learning Community (DMLC) and GB is a level-wise combination modelling. Initially, a weak classifier applies to the data. This further fits into another weaker classifier to enhance the performance of the existing model, without modifying the preceding classifier and then, the process continues so on. Every classifier would be noting that the preceding classifiers did not work best. In the First, x1 is estimated by adjusting the data into a decision tree. Secondly, the tree is built based on the remainder of the previous step of x-x1, and the process persists.

6.8 Results and Discussion

The enhanced IMO_DNN diabetes risk prediction model was proposed to tackle the issues encountered because of failing to select the best possible features in the last step. Likewise, the previous model, this enhanced model is also evaluated with the existing methodologies like RNN, CNN, and ANN in addition to the previously designed IMO_DNN model. The Accuracy, Recall, Precision and F1-score are being estimated by the deployment of the confusion matrix as given in chapter 3 (Eq. 3.1, 3.2, 3.3 and 3.4).

The training and testing for this comparative study is also similar to the previous phase of work, wherein 70% for training and 30% for testing is being done. The proposed enhanced IMO_DNN is being compared with the previously designed IMO_DNN model and other existing models as tabulated in the below table 1.

Table 6.1 Comparative study of Enhanced IMO_DNN

Evaluation Metrics	Accuracy	Precision	Recall	F1 score
EIMOXB_DNN	94.27	95.67	90.93	91.28
IMO_DNN	92.27	93.67	90.93	92.28
ANN	82.36	87.01	78.52	87.11
RNN	80.35	85.12	76.63	85.21
CNN	72.88	76.51	69.23	73.74

A graph is being plotted by taking the various models in the x-axis and its corresponding performance metrics in the y-axis.

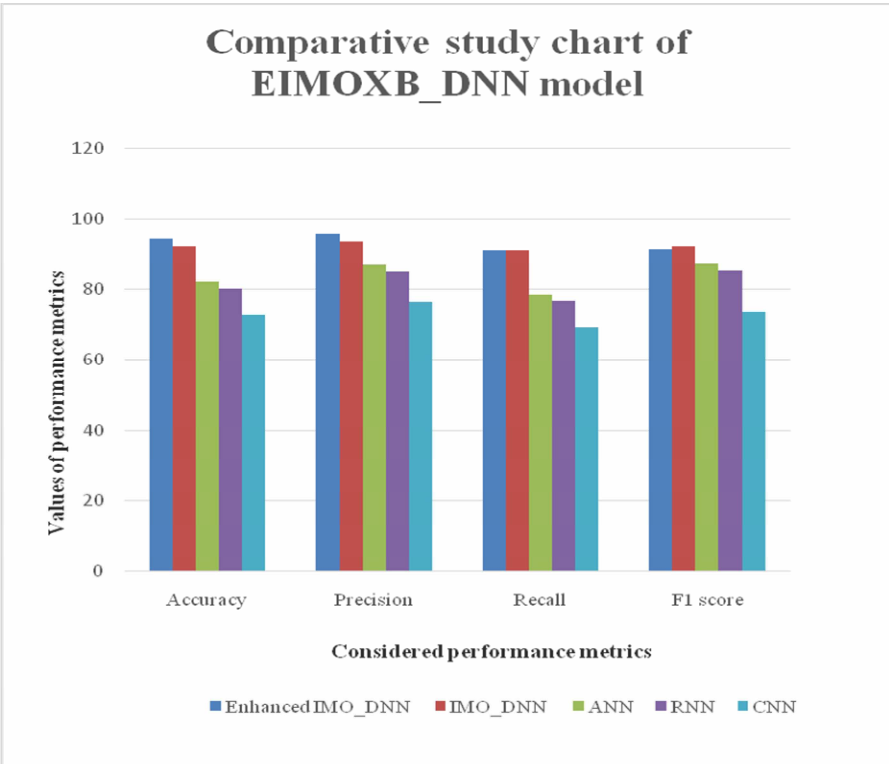


Fig. 6.1 Comparison of the proposed enhanced IMO_DNN with other models

From the graph, it could be identified that our proposed EIMOXB_DNN model was able to produce better of 94.27% than other models like ANN, RNN, and CNN with 82.36, 80.35, and 72.88 respectively, including the previous phase of IMO_DNN model with 92.27% of accuracy.

In terms of the precision, the proposed EIMOXB_DNN model was better with 95.67% over the lower precision values of 93.67, 85.12, 76.51, and 87.01 yielded by models like IMO_DNN, RNN, CNN, and ANN, respectively.

When taking the F1 score, a lower value 91.28 than the earlier IMO_DNN with 92.28 was produced by EIMOXB_DNN models, however, the F1 score was higher than those of the values yielded by other three models.

Finally, when taking the recall to consideration, both the EIMOXB_DNN and IMO_DNN models produced the value of 90.93%, which is anyway higher than the other three models.

From the comprehensive results comparison, the proposed model was able to outperform all other models considered including the previously proposed IMO_DNN in terms of accuracy and precision. However, the proposed model was not able to surpass the performance set by the previously proposed IMO_DNN model in terms of F1 score and recall.

6.9 Chapter Summary

In this current work, an EIMOXB_DNN model was devised to predict the risk of diabetes disease by eliminating the issues encountered if the selection of best features were not done. The proposed model was built in the same way as previously proposed IMO_DNN with the deep network that trained features by using the hidden layers. The processed ontology knowledge of collected data from earlier IMO_DNN had been given as an input. Finally, the output layer predicted the exact result like low risk or high risk or medium risk with better results. The experimental results of the

EIMOXB_DNN were compared with other few existing models and previously proposed IMO_DNN.

It is evident that the results of the enhanced IMO_DNN model outperformed all other models by yielding values of accuracy of 94.27%; precision of 94.27%; F1 score of 91.28; and recall of 90.93%. The currently proposed model predicted the risk of diabetes disease in an effective way in terms of both accuracy and precision.

Chapter – 7

Conclusion

CHAPTER - 7
CONCLUSION

7.1 Summary of Research

The currently pursued research work deduced improvised models of predicting and realizing the risks levels in Diabetes diseases. The research methodology and the methodology diagram along with the working of two phases were discussed in chapter 3. Chapter 3 also discussed the deployed data collection sources like IoT sensors, Questionnaires, and Clinical records. To get the domain-specific knowledge and insights, we have collected an enormous quantity of data to create our dataset for simulation and investigation. The successive chapters 4, 5 and 6 individually discussed the background, related works, working methodologies, algorithmic functions involved, and performance evaluation of three models.

- Rule Embedded Semantic Ontology-based DNN Classifier (RESOC_DNN) model.
- Improved-DNN with Binary Multi-Objective Optimization (IMO_DNN) Model for Predicting the Risk of Diabetes Patients.
- An Enhanced IMO_DNN (EIMOXB_DNN) Model with Booster to Improve the Accuracy of Risk Prediction.

The initial and first model considered for the processing of IoT data was based on the phenomenon of semantics, which processed the IoT data collected to generate ontology and rules. Afterwards, the DNN algorithm was applied to facilitate the process of training and testing to predict diabetes diseases. The next model attempted to categorize the risk levels of diabetes into three levels by using the appropriate loss function and binary multiple-objective optimization to predict the diabetes diseases. The third and final model was proposed as an improvement to the second model, which used the feature selection methodology by XGBoost along with the classification and Adam optimization to accurately predict diabetes along with its risk levels. However, the third model was only able to produce better results in terms of accuracy and precision. All these performance measures will be discussed in detail in the respective chapters.

7.2 Features of Proposed Models

7.2.1 RESOC_DNN Model

The Rule Embedded Semantic Ontology-based DNN Classifier operated based on three phases like Data extraction, modelling and result evaluation. The ontology data was prepared from the initially gathered data genres like Gender, Age, Glucose, Skin thickness, Insulin, BMI, Waistline, and Calorie intake, etc. Then the rule generation and DNN network classification takes place to predict diabetes diseases. The objectives achieved through this model were:

- Ontology data was prepared successfully developed with the deployment of IoT and domain data.
- The rule engine was developed to generate and validate the ontology.
- A DNN-based diabetes disease classification was made to predict the diseases.

7.2.2 IMO_DNN Model

In addition to the diabetes disease prediction, this model also categorized the risk levels into three, namely: Low risk, Medium risk and High risk. Two loss functions were used to ensure the closeness of the output towards the perfect value and binary multiple-objective optimization was applied to optimize the study parameters for facilitating the experimental validation. The objectives achieved through this model were:

- The accuracy of the collected dataset was enhanced than the previous model using DNN.
- The unsupervised data was successfully classified for the prepared dataset.
- The target class was predicted with improved accuracy.

7.2.3 EIMOXB_DNN Model

In addition to the diabetes disease prediction and risk level categorization, this model deployed the feature selection methodology to reduce the computational speed and time taken to predict diabetes disease. This model also made use of the loss and Adam optimization functions alongside the classification to optimize the results of the model. The objectives achieved through this model were:

- The hybrid model using the DNN and XGBoost was successfully integrated.
- The risk levels of diabetes disease were successfully predicted.
- The accuracy of this model was improved than the previous models.

7.3 Comparative study of Proposed Models

Performance metrics like accuracy, precision, recall, and F1 score were tabulated and graphs show the results of all the three models proposed in this research. The below table 7.1 details this comparative study.

Table 7.1 Comparative Results Investigation of all Proposed Models

Common Measures	RESOC- DNN	IMO_DNN	EIMOXB_DNN
Accuracy	88.33	92. 27	94. 27
Precision	88.67	93.67	95.67
Recall	87.3	90.93	90.93
F1 score	88.42	92.28	91.28

A graph showing the performance of the three models are depicted in the below fig. 7.1 with proposed models taken on the x-axis side and performance measure values taken in the y-axis (also in the z-axis).

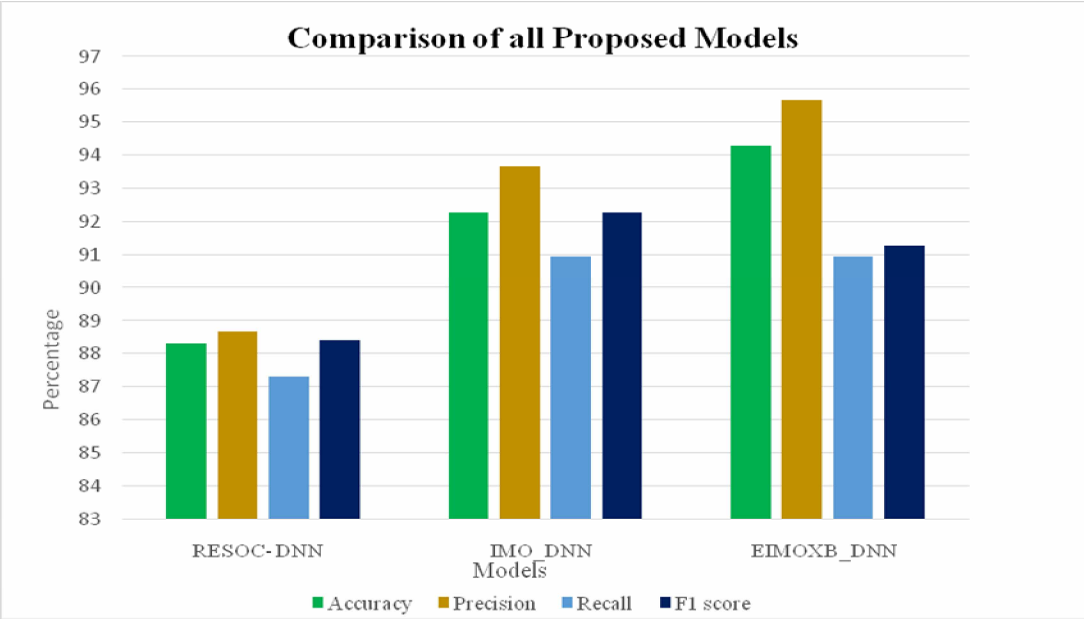


Fig. 7.1 Comparative Results Investigation of all Proposed Models

It could be inferred that the accuracy and precision values of 94.27% and 95.67% was yielded by the EIMOXB_DNN were higher than other two previously proposed models. Same recall values of 90.93% yielded by both the Enhanced IMO_DNN and IMO_DNN, lower than the RESOC_DNN model. The higher F1 score value of 92.28 was yielded by IMO_DNN surpassing the other two models.

7.4 Inferences from the Proposed Models

After the comprehensive investigation, it is inferred that all the three proposed models were yielding better results than the compared existing works of RNN, CNN, and ANN in terms of accuracy, precision, recall, and F1 score. The final enhanced IMO_DNN model was able to predict the risk level of diabetes disease by favorably selecting the features with improvised accuracy and precision than the two other proposed models like IMO_DNN and RESOC_DNN.

7.5 Drawbacks of the Proposed Models

All three models were proposed based on the IoT and domain data after collecting health-based data from three varied data sources. Even though all the proposed models gave rise to expect outcomes surpassing the existing methods, these models are limited only to the health care application and also lacks the capability of handling the imbalanced type of data if any.

7.6 Recommendations for Future Works

The final model EIMOXB_DNN was able to provide the improvised risk prediction than other proposed methods. Furthermore advanced feature selection methodology could be adopted to put forth an even more improvised diabetes disease prediction. Since all the proposed models were primarily based on IoT semantics, more diverse Web semantic technologies could be used and the range of data collection could also be further extended up to the state level rather than collecting over a few patients in a specific region.

References

REFERENCES

[Abi, 18] Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat Abd Elatif Mohamed and Humaira Arshad, "State-of-the-art in artificial neural network applications: A survey", Heliyon, Vol. 4, No. 11, pp. 1-41, 2018.

[Ade, 14] Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi and Charles Korede Ayo, "Comparison of ARIMA and artificial neural networks models for stock price prediction", Journal of Applied Mathematics, pp. 1-8, 2014. DOI: https://doi.org/10.1155/2014/614342.

[Ahl, 20] Ahlem Rhayem, Mohamed Ben Ahmed Mhiri and Faiez Gargouri, "Semantic Web Technologies for the Internet of Things: Systematic Literature Review", Internet of Things, Vol. 11, ISSN: 2542-6605, pp. 1-22, 2020.

[Ahm, 21] Ahmad, Hafiz Farooq, Hamid Mukhtar, Hesham Alaqail, Mohamed Seliaman and Abdulaziz Alhumam, "Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning", Applied Sciences, Vol.11, No. 3, pp. 1-18, 2021.

[AIZ, 20] AIZubi and Ahmad Ali, "Big data analytic diabetics using map reduce and classification techniques", The Journal of Supercomputing, Vol. 76, No. 6, pp. 4328-4337, 2020.

[Aja, 18] Ajami, Hicham and Hamid Mcheick, "Ontology-based model to support ubiquitous healthcare systems for COPD patients", Electronics, Vol. 7, No. 12, pp. 1-29, 2018.

[Ala, 17] Alansari Z., Anuar N. B., Kamsin A., Soomro S. and Belgaum, M. R, "The Internet of Things adoption in healthcare applications", International Conference on Engineering Technologies and Social Sciences (ICETSS), pp.1-5, 2017. DOI: 10.1109/ICETSS.2017.8324138.

[Ala, 18] Alamri and Abdullah, "Ontology middleware for integration of IoT healthcare information systems in EHR systems", Computers, Vol. 7, No. 4, pp.1-15, 2018.

[Alh, 19] Alhakbani, Noura, Mohammad Mehedi Hassan, Mourad Ykhlef and Giancarlo Fortino, "An efficient event matching system for semantic smart data in the Internet of Things (IoT) environment", Future Generation Computer Systems, Vol. 95, ISSN: 0167-739X, pp.163-174, 2019.

[Ali, 17] Alic, Berina, Lejla Gurbeta and Almir Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases", Mediterranean conference on embedded computing (MECO), pp. 1-4, 2017. DOI: 10.1109/MECO.2017.7977152.

[Ali, 18a] Ali, Farman, SM Riazul Islam, Daehan Kwak, Pervez Khan, Niamat Ullah, Sang-jo Yoo and Kyung Sup Kwak, "Type-2 fuzzy ontology–aided recommendation systems for IoT-based healthcare", Computer Communications, Vol. 119, ISSN: 0140-3664, pp. 138-155, 2018.

[Ali, 18b] Ali, Mohammed Hasan, Bahaa Abbas Dawood Al Mohammed, Alyani Ismail and Mohamad Fadli Zolkipli, "A new intrusion detection system based on fast learning network and particle swarm optimization", IEEE Access, Vol. 6, pp. 20255-20261, 2018. DOI. 10.1109/ ACCESS. 2018.2820092.

[And, 18] Androcec, Darko, Matija Novak and Dijana Oreski, "Using semantic web for internet of things interoperability: A systematic review", International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 14, No. 4, pp. 147-171, 2018.

[Ara, 17] Ara, Affreen and Aftab Ara, "Case study: Integrating IoT, streaming analytics and machine learning to improve intelligent diabetes management system", International conference on energy, communication, data analytics and soft computing (ICECDS), pp. 3179-3182, 2017. DOI: 10.1109/ICECDS.2017.8390043.

[Aro, 19a] Arockiam L., Sathyapriya S., Jane V. A. and Dalvin Vinoth Kumar A., "Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India", International Journal of Recent Technology and Engineering, Vol. 8, No. 2, pp. 6440-6444, 2019.

[Aro, 19b] Arockiam L., Sathyapriya S., Jane V. A. and Dalvin Vinoth Kumar A., "Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning", IJRTE Blue Eyes Intelligence Engineering & Sciences Publication, Vol. 8, No. 2, pp. 6400-6403, 2019.

[Asg, 19] Asghari, Parvaneh, Amir Masoud Rahmani and Hamid Haj Seyyed Javadi, "Internet of Things applications: A systematic review", Computer Networks, Vol.148, ISSN: 1389-1286, pp. 241-261, 2019.

[Asi, 18] Asim, Muhammad Nabeel, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood and Hafiza Mahnoor Abbasi, "A survey of ontology learning techniques and applications", Database, pp.1-24, 2018.DOI: https://doi.org/10.1093/database/bay101.

[Ati, 20] Atitallah, Safa Ben, Maha Driss, Wadii Boulila and Henda Ben Ghezala, "Leveraging Deep Learning and IoT big data analytics to support the smart cities development: Review and future directions", Computer Science Review, Vol. 38, ISSN: 1574-0137, pp.1-29, 2020.

[Baj, 17] Bajaj, Garvita, Rachit Agarwal, Pushpendra Singh, Nikolaos Georgantas and Valerie Issarny, "A study of existing Ontologies in the IoT-domain", pp.1-24, 2017. DOI: https://doi.org/10.48550/arXiv.1707.00112.

[Ban, 18] Banane, Mouad and Abdessamad Belangour, "A Survey on RDF Data Store Based on NoSQL Systems for the Semantic Web Applications", International Conference on Advanced Intelligent Systems for Sustainable Development, pp. 444-451, 2018. DOI: http://dx.doi.org/ 10.1007/978-3-030-11928-7_40.

[Baq, 19] Baqa, Hamza, Martin Bauer, Sonia Bilbao, Aitor Corchero, Laura Daniele, Iker Esnaola and Izaskun Fernández, "Semantic IoT Solutions - A Developer Perspective", pp. 1-51, 2019. DOI: http://dx.doi.org/ 10.13140/RG.2.2.16339.53286.

[Bar, 12] Barnaghi P., Wang W., Henson C., Taylor K. and Systems I., "Semantics for the Internet of Things: early progress and back to the future, International journal on Semantic Web and information systems, Vol. 8, No. 1, pp. 1-21, 2012.

[Bas, 19] Bashar and Abul, "Survey on evolving deep learning neural network architectures", Journal of Artificial Intelligence, Vol. 1, No. 02, pp. 73-82, 2019.

[Bat, 19] Batbaatar, Erdenebileg and Keun Ho Ryu, "Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach", International journal of environmental research and public health, Vol. 16, No. 19, pp. 1-19, 2019.

[Bee, 20] Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk and Laura M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy", CHI Conference on Human Factors in Computing Systems, pp. 1-12, 2020, DOI: https://doi.org/10.1145/3313831.3376718.

[Ber, 12] Bergstra, James and Yoshua Bengio, "Random search for hyper-parameter optimization", Journal of machine learning research, Vol. 13, No. 2, pp. 281-305, 2012.

[Ber, 17] Bermudez-Edo, Maria, Tarek Elsaleh, Payam Barnaghi and Kerry Taylor, "IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics", Personal and Ubiquitous Computing, Vol. 21, No. 3, pp. 475-487, 2017.

[Bha, 20] Bhatia, Munish, Simranpreet Kaur, Sandeep K. Sood and Veerawali Behal, "Internet of things-inspired healthcare system for urine-based diabetes prediction", Artificial Intelligence in Medicine, Vol. 107, ISSN: 0933-3657, pp. 1-16, 2020.

[Bis, 20] Biswas, Riya, Souvik Pal, Bikramjit Sarkar and Arindam Chakrabarty, "Health-Care Paradigm and Classification in IoT Ecosystem Using Big Data Analytics: An Analytical Survey", Intelligent Computing in Engineering: Select Proceedings of RICE, 2020. DOI: https://doi.org/10.1016/C2019-0-00358-5.

[Bur, 19] Burzlaff, Fabian, Nils Wilken, Christian Bartelt and Heiner Stuckenschmidt, "Semantic interoperability methods for smart service systems: A survey", IEEE Transactions on Engineering Management, pp. 1-15, 2019. DOI: 0.1109/TEM.2019.2922103.

[Cal, 18] Caliskan A., Yuksel ME., Badem H. and Basturk A., "Performance improvement of deep neural network classifiers by a simple training strategy", Eng Appl Artif Intell, Vol. 67, ISSN: 0952-1976, pp. 14–23, 2018.

[Can, 18] Cano, Juan Carlos, Victor Berrios, Ben Garcia and Chai K. Toh, "Evolution of IoT: an industry perspective", IEEE Internet of Things Magazine, Vol. 1, No. 2, pp. 12-17, 2018.

[Che, 16] Chen T. and Guestrin C., "XGBoost: a scalable tree boosting system", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016. DOI: https://doi.org/10.1145/2939672.2939785.

[Che, 18] Chen, Zhuo, Fu Jiang, Yijun Cheng, Xin Gu, Weirong Liu and Jun Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud", IEEE international conference on big data and smart computing (bigcomp), pp. 251-256, 2018. DOI: 10.1109/BigComp.2018.00044.

[Che, 19] Chen, Li, Dongxin Lu, Menghao Zhu, Muhammad Muzammal, Oluwarotimi Williams Samuel, Guixin Huang, Weinan Li and Hongyan Wu, "OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems", International Journal of Distributed Sensor Networks, Vol. 15, No. 5, pp. 1-15, 2019.

[Che, 20] Chen, Guang, Tonghai Jiang, Meng Wang, Xinyu Tang and Wenfei Ji, "Modeling and reasoning of IoT architecture in semantic ontology dimension", Computer Communications, Vol. 153, pp. 580-594, 2020. DOI: http://dx.doi.org/10.1016/j.comcom.2020.02.006.

[Cho, 18] Choo, Jaegul and Shixia Liu, "Visual analytics for explainable deep learning", IEEE computer graphics and applications, Vol. 38, No. 4, pp. 84-92, 2018.

[Da, 14] Da Xu, Li, Wu He and Shancang Li, "Internet of things in industries: A survey", IEEE Transactions on industrial informatics, Vol. 10, No. 4, pp. 2233-2243, 2014.

[Den, 12] Deng Li, "Three classes of deep learning architectures and their applications: a tutorial survey", APSIPA transactions on signal and information processing, pp. 1-28, 2012.

[Dha, 18] Dhaliwal, Sukhpreet Singh, Abdullah-Al Nahid and Robert Abbas, "Effective intrusion detection system using XGBoost", Information, Vol. 9, No. 7, pp. 1-24, 2018.

[Dis, 20] Dissanayake, Pavithra I., Tiago K. Colicchio and James J. Cimino, "Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis", Journal of the American Medical Informatics Association, Vol. 27, No. 1, pp. 159-174, 2020.

[Dje, 20] Djerioui, Mohamed, Youcef Brik, Mohamed Ladjal and Bilal Attallah, "Heart Disease prediction using MLP and LSTM models", International Conference on Electrical Engineering (ICEE), pp. 1-5, 2020. DOI: 10.1109/ICEE49691.2020.9249935.

[Dud, 18] Dudas, Marek, Steffen Lohmann, Vojtech Svatek and Dmitry Pavlov, "Ontology visualization methods and tools: A survey of the state of the art", The Knowledge Engineering Review, Vol. 33, pp. 1-39, 2018. DOI: http://dx.doi.org/10.1017/S0269888918000073.

[Est, 19] Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark De Pristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun and Jeff Dean, “A guide to deep learning in healthcare”, Nature medicine, Vol. 25, No. 1, pp. 24-29, 2019.

[Far, 18] Farahani, Bahar, Farshad Firouzi, Victor Chang, Mustafa Badaroglu, Nicholas Constant and Kunal Mankodiya, “Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare”, Future Generation Computer Systems, Vol. 78, pp. 659-676, 2018. DOI: https://doi.org/10.1016/j.future.2017.04.036.

[Gan, 17] Ganzha, Maria, Marcin Paprzycki, Wiesław Pawłowski, Paweł Szmeja and Katarzyna Wasielewska, “Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective”, Journal of Network and Computer Applications, Vol. 81, pp. 111-124, 2017. DOI: http://dx.doi.org/10.1016/j.jnca.2016.08.007.

[Gha, 19] Ghani, Norjihan Abdul, Suraya Hamid, Ibrahim Abaker Targio Hashem and Ejaz Ahmed, “Social media big data analytics: A survey”, Computers in Human Behavior, Vol. 101, pp. 417-428, 2019. DOI: https://doi.org/10.1016/j.chb.2018.08.039.

[Goe, 17] Goel, Deepti, Nisha Pahal, Parul Jain and Santanu Chaudhury, “An ontology-driven context aware framework for smart traffic monitoring”, IEEE Region 10 Symposium (TENSYMP), pp. 1-5, 2017. DOI: 10.1109/ TENCONSpring.2017.8070059.

[Goh, 20] Gohel, Hardik A., Himanshu Upadhyay, Leonel Lagos, Kevin Cooper and Andrew Sanzetenea, “Predictive maintenance architecture development for nuclear infrastructure using machine learning”, Nuclear Engineering and Technology, Vol. 52, No. 7, pp. 1436-1442, 2020.

[Gou, 18] Guo, Yiqing, Xiuping Jia and David Paull, “Effective sequential classifier training for SVM-based multitemporal remote sensing image classification”, IEEE Transactions on Image Processing, Vol. 27, No. 6, pp. 3036-3048 , 2018.

[Gru, 18] Gruber Tom, “Ontology”, pp. 1-11, 2018. Accessed on 13, December 2021.

[Gub, 13] Gubbi, Jayavardhana, Rajkumar Buyya, Slaven Marusic and Marimuthu Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions”, Future generation computer systems, Vol. 29, No. 7, pp. 1645-1660, 2013.

[Gue, 20] Gue, Ivan Henderson V., Aristotle T. Ubando, Ming-Lang Tseng and Raymond R. Tan, “Artificial neural networks for sustainable development: a critical review”, Clean Technologies and Environmental Policy, Vol. 22, pp. 1449-1465, 2020. DOI: https://doi.org/10.1007/s10098-020-01883-2.

[Gup, 21] Gupta, Sakshi and Umang Singh, “Ontology-Based IoT Healthcare Systems (IHS) for Senior Citizens”, International Journal of Big Data and Analytics in Healthcare (IJBDAH), Vol. 6, No. 2 , pp. 1-17, 2021.

[Gyr, 18a] Gyrard, Amelie, Antoine Zimmermann and Amit Sheth, “Building IoT-based applications for smart cities: How can ontology catalogs help”, IEEE Internet of Things Journal, Vol. 5, No. 5, pp. 3978-3990, 2018.

[Gyr, 18b] Gyrard, Amelie, Soumya Kanti Datta and Christian Bonnet, “A survey and analysis of ontology-based software tools for semantic interoperability in IoT and WoT landscapes”, World Forum on Internet of Things (WF-IoT), pp. 86-91, 2018. DOI: 10.1109/WF-IoT.2018.8355091.

[Hal, 17] Haller, Armin, Krzysztof Janowicz, Simon Cox, Danh Le Phuoc, Kerry Taylor, Maxime Lefrancois, R. Atkinson, R. Garcıa-Castro, J. Lieberman and C. Stadler, “Semantic Sensor Network Ontology”, W3C Recommendation, pp. 1-33, 2017.

[Haq, 20] Haq, Amin Ul, Jianping Li, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad and Amjad Ali, “A New Intelligent Approach for Effective Recognition of Diabetes in the IoT E-HealthCare Environment”, pp. 1-30, 2020. DOI: 10.20944/preprints202002.0462.v1.

[Har, 14] Harrison, Robert, C. Stuart McLeod, Giacomo Tavola, Marco Taisch, Armando W. Colombo, Stamatis Karnouskos and Marcel Tilly, "Next generation of engineering methods and tools for SOA-based large-scale and distributed process applications", In Industrial cloud-based cyber-physical systems, pp. 137-165, 2014. DOI: 10.1007/978-3-319-05624-1_6.

[Har, 21] Harimoorthy, Karthikeyan and Menakadevi Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system", Journal of Ambient Intelligence and Humanized Computing Vol.12, No. 3, pp.3715-3723, 2021.

[Has, 20] Hasan, Md Kamrul, Md Ashraful Alam, Dola Das, Eklas Hossain and Mahmudul Hasan, "Diabetes prediction using ensembling of different machine learning classifiers", IEEE Access, Vol. 8, pp. 76516-76531, 2020. DOI: 10.1109/ACCESS.2020.2989857.

[Hin, 06] Hinton, Geoffrey E., Simon Osindero and Yee-Whye The, "A fast learning algorithm for deep belief nets", Neural computation, Vol. 18, No. 7, pp. 1527-1554, 2006.

[Hoh, 20] Hohenecker, Patrick and Thomas Lukasiewicz, "Ontology reasoning with deep neural networks", Journal of Artificial Intelligence Research, Vol. 68, pp. 5060-5064, 2020. DOI: https://doi.org/10.24963/ ijcai.2020/707.

[Hon, 19] Honti, Gergely Marcell and Janos Abonyi, "A review of semantic sensor technologies in internet of things architectures", Complexity, pp. 1-22, 2019. DOI: https://doi.org/10.1155/2019/6473160.

[IBM, 21] IBM, https://www.ibm.com/cloud/learn/deep-learning, 2021. Accessed on 13, December 2021.

[Inf, 21] Influential Points, https://influentialpoints.com/Training/analytical_surveys – principles – properties - assumptions.htm, 2021.

[Isl, 20] Islam, Md Merajul, Md Jahanur Rahman, Dulal Chandra Roy and Md Maniruzzaman, "Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach", Diabetes & Metabolic Syndrome: Clinical Research & Reviews, Vol. 14, No. 3, pp. 217-219, 2020.

[Jab, 17] Jabbar, Sohail, Farhan Ullah, Shehzad Khalid, Murad Khan and Kijun Han, "Semantic interoperability in heterogeneous IoT infrastructure for healthcare", Wireless Communications and Mobile Computing, pp. 1-10, 2017. DOI: https://doi.org/10.1155/2017/9731806.

[Jen, 17] Jennex M., "Big data, the internet of things, and the revised knowledge pyramid", Data Base for Advances in Information Systems, Vol. 48, No. 4, pp. 69-79, 2017.

[Jia, 17a] Jia, Bing, Wuyungerile Li and Tao Zhou, "A centralized service discovery algorithm via multi-stage semantic service matching in internet of things", IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Vol. 1, pp. 422-427, 2017. DOI: 10.1109/CSE-EUC.2017.82.

[Jia, 17b] Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen and Yongjun Wang, "Artificial intelligence in healthcare: past, present and future", Stroke and vascular neurology, Vol. 2, No. 4, pp. 1-14, 2017.

[Kam, 19] Kamakshi Priyaa P., Sathyapriya S. and Arockiam L., "Nutrition Monitoring and Calorie Estimation using Internet of Things (IoT)", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 11, pp. 2669-2672, 2019.

[Kar, 15] Karagiannis, Vasileios, Periklis Chatzimisios, Francisco Vazquez-Gallego and Jesus Alonso-Zarate, "A survey on application layer protocols for the internet of things", Transaction on IoT and Cloud computing, Vol. 3, No. 1, pp. 11-17, 2015.

[Kio, 19a] Kiourtis, Athanasios, Argyro Mavrogiorgou, Andreas Menychtas, Ilias Maglogiannis and Dimosthenis Kyriazis, “Structurally mapping healthcare data to HL7 FHIR through ontology alignment”, Journal of Medical Systems, Vol. 43, No. 3, pp. 1-13, 2019.

[Kio, 19b] Kiourtis, Athanasios, Sokratis Nifakos, Argyro Mavrogiorgou and Dimosthenis Kyriazis, “Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching”, International Journal of Medical Informatics, Vol. 132, 2019. DOI: http://dx.doi.org/10.1016/ j.ijmedinf.2019.104002.

[Kop, 20] Kopitar, Leon, Primoz Kocbek, Leona Cilar, Aziz Sheikh and Gregor Stiglic, “Early detection of type 2 diabetes mellitus using machine learning-based prediction models”, Scientific reports, Vol. 10, No. 1, pp. 1-12, 2020.

[Kot, 12] Kotis K. and Katasonov A., VTT Technical Research Center, Finland VTT Technical Research Center, Finland, “An iot-ontology for the representation of interconnected, clustered and aligned smart entities”, VTT Technical Research Center, Finland VTT Technical Research Center, Finland, pp. 1-14, 2012.

[Kri, 20] Krishnamurthi, Rajalakshmi, Adarsh Kumar, Dhanalekshmi Gopinathan, Anand Nayyar and Basit Qureshi, “An overview of IoT sensor data processing, fusion, and analysis techniques”, Sensors, Vol. 20, No. 21, pp. 1-23, 2020.

[Lal, 20] Lal, Devesh Kumar and Ugrasen Suman, “A Survey of Real-Time Big Data Processing Algorithms”, In Reliability and Risk Assessment in Engineering, pp. 3-10, 2020. DOI: 10.1186/s13634-016-0355-x.

[Le, 20] Le, Tuan Minh, Thanh Minh Vo, Tan Nhat Pham and Son Vu Truong Dao, “A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic”, IEEE Access, Vol. 9, pp. 7869-7884, 2020. DOI: 10.1109/ACCESS.2020.3047942.

[Li, 20] Li, Shenglong and Xiaojing Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm", Neural Computing and Applications, Vol. 32, No. 7, pp. 1971-1979, 2020.

[Li, 21] Li, Yanxia, Yi Chai, Han Zhou and Hongpeng Yin, "A novel dimension reduction and dictionary learning framework for high-dimensional data classification", Pattern Recognition, Vol. 112, ISSN: 0031-3203, pp. 1-43, 2021.

[Lia, 18] Liang, Fan, Wei Yu, Dou An, Qingyu Yang, Xinwen Fu and Wei Zhao, "A survey on big data market: Pricing, trading and protection", IEEE Access, Vol. 6, pp. 15132-15154, 2018. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2806881.

[Liu, 17] Liu, Weibo, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu and Fuad E. Alsaadi, "A survey of deep neural network architectures and their applications", Neurocomputing, Vol. 234, pp. 11-26, 2017. DOI: https://doi.org/10.1016/j.neucom.2016.12.038.

[Lou, 18] Loukil, Faiza, Chirine Ghedira-Guegan, Khouloud Boukadi and Aicha Nabila Benharkat, "LIoPY: A legal compliant ontology to preserve privacy for the Internet of Things", Annual Computer Software and Applications Conference (COMPSAC), Vol. 2, pp. 701-706, 2018. DOI: http://dx.doi.org/10.1109/COMPSAC.2018.10322.

[Ma, 19] Ma, Xiaoqiang, Tai Yao, Menglan Hu, Yan Dong, Wei Liu, Fangxin Wang and Jiangchuan Liu, "A survey on deep learning empowered IoT applications", IEEE Access, Vol. 7, pp. 181721-181732, 2019.

[Man, 15] Mantovani, Rafael G., Andre LD Rossi, Joaquin Vanschoren, Bernd Bischl and Andre CPLF De Carvalho, "Effectiveness of random search in SVM hyper-parameter tuning", International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2015. DOI: 10.1109/ IJCNN.2015.7280664.

[Man, 20a] Maniruzzaman Md, Md Jahanur Rahman, Benojir Ahammed and Md Menhazul Abedin, "Classification and prediction of diabetes disease using machine learning paradigm", Health information science and systems, Vol. 8, No. 1, pp. 1-14, 2020.

[Man, 20b] Manonmani M. and Sarojini Balakrishnan, "A review of semantic annotation models for analysis of healthcare data based on data mining techniques", Emerging Research in Data Engineering Systems and Computer Communications, pp. 231-238, 2020. DOI: https://doi.org/10.1186/2196-1115-1-2.

[Mar, 14] Martin-Timon, Iciar, Cristina Sevillano-Collantes, Amparo Segura-Galindo and Francisco Javier del Canizo-Gomez, "Type 2 diabetes and cardiovascular disease: have all risk factors the same strength", World journal of diabetes, Vol. 5, No. 4, pp. 444-470, 2014.

[Mar, 18] Martinez-Caro, Eva, Juan Gabriel Cegarra-Navarro, Alexeis Garcia-Perez and Monica Fait, "Healthcare service evolution towards the Internet of Things: An end-user perspective", Technological Forecasting and Social Change, Vol. 136, pp. 268-276, 2018. DOI: https://doi.org/10.1016/j.techfore.2018.03.025.

[Mav, 19] Mavrogiorgou, Argyro, Athanasios Kiourtis, Konstantinos Perakis, Stamatios Pitsios and Dimosthenis Kyriazis, "IoT in healthcare: Achieving interoperability of high-quality data acquired by IoT medical devices", Sensors, Vol. 19, No. 9, pp. 1-24, 2019.

[Mer, 21] Mervin R., Tintu Thomas and Jaya A., "Ontology-Based System for Patient Monitoring", Semantic Web for Effective Healthcare, pp. 75-93, 2021. DOI: http://dx.doi.org/10.1002/9781119764175.ch3.

[Mir, 18] Mironczuk, Marcin Michal and Jaroslaw Protasiewicz, "A recent overview of the state-of-the-art elements of text classification", Expert Systems with Applications, Vol. 106, pp. 36-54, 2018. DOI: https://doi.org/10.1016/j.eswa.2018.03.058.

[Mit, 19] Mittal S. and Sangwan OP., "Big data analytics using machine learning techniques", International Conference on Cloud Computing, Data Science & Engineering, pp. 203-207, 2019. DOI: 10.1109/CONFLUENCE. 2019.8776614.

[Moh, 18] Mohammadi, Mehdi, Ala Al-Fuqaha, Sameh Sorour and Mohsen Guizani, "Deep learning for IoT big data and streaming analytics: A survey", IEEE Communications Surveys & Tutorials, Vol. 20, No. 4, pp. 2923-2960, 2018.

[Moh, 21] Mohammed Ali A., Mohammed A. Noaman Al-hayanni and Hassan M. Azzawi, "Detection and segmentation the affected brain using ThingSpeak platform based on IoT cloud analysis", Periodicals of Engineering and Natural Sciences (PEN), Vol. 9, No. 2, pp. 858-867, 2021.

[Mor, 18] Moreira, Joao, Luis Ferreira Pires, Marten van Sinderen and Laura Daniele, "SAREF4health: IoT Standard-Based Ontology-Driven Healthcare Systems", In FOIS, pp. 239-252, 2018. DOI: http://dx.doi.org/10.3233/ 978-1-61499-910-2-239.

[Mou, 16] Moustafa, Hassnaa, Eve M. Schooler, Gang Shen and Sanjana Kamath, "Remote monitoring and medical devices control in eHealth", International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1-8, 2016. DOI: 10.1109/ WiMOB.2016.7763177.

[Nag, 21] Nagowah, Soulakshmee D., Hatem Ben Sta and Baby Gobin-Rahimbux, "A systematic literature review on semantic models for IoT-enabled smart campus", Applied Ontology Preprint, pp. 1-27, 2021. DOI: 10.3233/ AO-200240.

[Naj, 21] Najafizadeh, Abbas, Afshin Salajegheh, Amir Masoud Rahmani and Amir Sahafi, "Privacy-preserving for the internet of things in multi-objective task scheduling in cloud-fog computing using goal programming approach", Peer-to-Peer Networking and Applications, Vol. 14, No. 6, pp. 3865-3890, 2021.

[Nav, 19] Naveed, Arjmand, Yim Fun Hu, Tshiamo Sigwele, Ghulam Mohi-Ud-Din and Misfa Susanto, "Similarity Analyzer for Semantic Interoperability of Electronic Health Records Using Artificial Intelligence (AI)", Journal of Engineering and Scientific Research, Vol. 1, No. 2, pp. 53-58, 2019.

[Nor, 19] Nord, Jeretta Horn, Alex Koohang and Joanna Paliszkiewicz, "The Internet of Things: Review and theoretical framework", Expert Systems with Applications, Vol. 133, pp. 97-108, 2019.

[Ole, 13] O'Leary and Daniel E, "Big Data', the 'Internet of Things' and the 'Internet of Signs", Intelligent Systems in Accounting, Finance and Management, Vol. 20, No. 1, pp. 53-65, 2013.

[Pal, 16] Palavalli, Amarnath, Durgaprasad Karri and Swarnalatha Pasupuleti, "Semantic internet of things", International Conference on Semantic Computing (ICSC), pp. 91-95, 2016. DOI: 10.1109/ICSC.2016.35.

[Pan, 21] Pandit, Arushi and Anchal Garg, "Artificial Neural Networks in Healthcare: A Systematic Review", International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 1-6, 2021. DOI: 10.1109/Confluence51648.2021.9377086.

[Par, 19] Park, Jinhyun, Seong-Jin Han, Nauman Munir, Yun-Taek Yeom, Sung-Jin Song, Hak-Joon Kim and Se-Gon Kwon, "MRPC eddy current flaw classification in tubes using deep neural networks", Nuclear Engineering and Technology, Vol. 51, No. 7, pp. 1784-1790, 2019.

[Pat, 16] Patel, Keyur K. and Sunil M. Patel, "Internet of things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges", International Journal of Engineering Science and Computing, Vol. 6, No. 5, pp. 6122 -6131, 2016.

[Pat, 18] Patel, Harsh H. and Purvi Prajapati, "Study and analysis of decision tree based classification algorithms", International Journal of Computer Sciences and Engineering, Vol. 6, No. 10, pp. 74-78, 2018.

[Pra, 18] Pradeep K. R. and Naveen N. C., "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4. 5 and Naive Bayes algorithms for healthcare analytics", Procedia computer science, Vol.132, pp. 412-420, 2018. DOI: https://doi.org/10.1016/j.procs.2018.05.162.

[Pur, 18] Purushotham, Sanjay, Chuizheng Meng, Zhengping Che and Yan Liu, "Benchmarking deep learning models on large healthcare datasets", Journal of biomedical informatics, Vol. 83, pp. 112-134, 2018.

[Rah, 19] Rahman, Md Abdur, Mamunur Rashid, Stuart Barnes, M. Shamim Hossain, Elham Hassanain and Mohsen Guizani, "An IoT and blockchain-based multi-sensory in-home quality of life framework for cancer patients", International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 2116-2121. 2019. DOI: 10.1109/IWCMC.2019.8766496.

[Rah, 20] Rahman, Hafizur and MdIftekhar Hussain, "A comprehensive survey on semantic interoperability for Internet of Things: State-of-the-art and research challenges", Transactions on Emerging Telecommunications Technologies, Vol. 31, No. 12, pp. 1-25, 2020.

[Raj, 20] Rajan, J. Pandia, S. Edward Rajan, Roshan Joy Martis and Bijaya K. Panigrahi, "Fog computing employed computer aided cancer classification system using deep neural network in internet of things based healthcare system", Journal of Medical Systems, Vol. 44, No. 2, pp. 1-10, 2020.

[Rat, 19] Rath M. and Pattanayak B., "Technological improvement in modern health care applications using Internet of Things (IoT) and proposal of novel health care approach", International Journal of Human Rights in Healthcare, Vol. 12, No. 2, pp. 148-162, 2019.

[Ray, 16] Ray and Partha Pratim, "A survey of IoT cloud platforms", Future Computing and Informatics Journal, Vol. 1, No. 1-2, pp. 35-46, 2016.

[Raz, 19] Raza, Muhammad Ahsan, Rahmah Mokhtar, Noraziah Ahmad, Maruf Pasha and Urooj Pasha, "A taxonomy and survey of semantic approaches for query expansion", IEEE Access, Vol. 7, pp. 17823-17833, 2019. DOI: 10.1109/ACCESS.2019.2894679.

[Reb, 18] Rebelo Moreira, Joao Luiz, Luis Ferreira Pires, Patricia Dockhorn Costa and Marten Van Sinderen, "Semantic Interoperability of Early Warning Systems: a Systematic Literature Review", Enterprise Interoperability: Smart Services and Business Impact of Enterprise Interoperability, pp. 407-413, 2018. DOI: http://dx.doi.org/10.1002/9781119564034.ch50.

[Reh, 19] Rehman, Muhammad Habib, Ibrar Yaqoob, Khaled Salah, Muhammad Imran, Prem Prakash Jayaraman and Charith Perera, "The role of big data analytics in industrial Internet of Things", Future Generation Computer Systems, Vol. 99, pp. 247-259, 2019. DOI: https://doi.org/10.1016/j.future.2019.04.020.

[Rom, 18] Roman-Castro, Rodrigo, Javier Lopez and Stefanos Gritzalis, "Evolution and trends in IoT security", Computer, Vol. 51, No. 7, pp. 16-25, 2018.

[Run, 20] Runkler and Thomas A., "Data analytics", Springer Fachmedien Wiesbaden, pp. 1-167, 2020. DOI: https://doi.org/10.1007/978-3-658-29779-4.

[Sah, 20] Sahoo, Abhaya Kumar, Chittaranjan Pradhan and Himansu Das, "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making", Nature inspired computing for data science, pp. 201-212, 2020. DOI: http://dx.doi.org/10.1007/978-3-030-33820-6_8.

[Sal, 21] Saleem, Tausifa Jan and Mohammad Ahsan Chishti, "Big Data Analytics for the Internet of Things: An Overview", Big Data Analytics for Internet of Things, pp. 1-5, 2021. DOI: https://doi.org/10.1002/9781119740780.ch1.

[Sam, 16] Samie, Farzad, Lars Bauer and Jorg Henkel, “IoT technologies for embedded computing: A survey”, International Conference on Hardware/Software Codesign and System Synthesis, pp. 1-10, 2016.

[San, 20] Sandhiya S. and Palani U., “An effective disease prediction system using incremental feature selection and temporal convolutional neural network”, Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 11, pp. 5547-5560, 2020.

[Sat, 20a] Sathyapriya S. and Arockiam L., “A Review On semantic Technology in IoT Healthcare”, Annals of the Romanian Society for Cell Biology, 2021, Vol. 25, No. 4, pp. 9810-9819, 2020.

[Sat, 20b] Sathyapriya S. and Arockiam L., “Rule Embedded Semantic Ontology Based Classifier For IoT Healthcare”, Annals of the Romanian Society for Cell Biology, 2021, Vol. 25, No. 4, pp. 10224-10231, 2020.

[Sca, 17] Scarpato, Noemi, Alessandra Pieroni, Luca Di Nunzio and Francesca Fallucchi, “E-health-IoT universe: A review”, Management, Vol. 21, No. 44, pp. 2328-2336, 2017.

[Sch, 15] Schmidhuber, Jurgen, “Deep learning in neural networks: An overview”, Neural Networks, Vol.61, pp. 85-117, 2015. DOI: https://doi.org/10.1016/j.neunet.2014.09.003.

[Sel, 10] Selvin, Elizabeth, Michael W. Steffes, Hong Zhu, Kunihiro Matsushita, Lynne Wagenknecht, James Pankow, Josef Coresh and Frederick L. Brancati, “Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults”, New England Journal of Medicine, Vol. 362, No. 9, pp. 800-811, 2010.

[Seo, 19] Seok, Hyun-Seung and Yong-Ju Lee, “Ontology-based IoT context information modeling and semantic-based IoT mashup services implementation”, The Journal of the Korea institute of electronic communication sciences, Vol. 14, No. 4, pp. 671-678, 2019.

[Sha, 20] Shafique, Kinza, Bilal A. Khawaja, Farah Sabir, Sameer Qazi and Muhammad Mustaqim, “Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios”, IEEE Access, Vol. 8, pp. 23022-23040, 2020. DOI: 10.1109/ACCESS.2020.2970118.

[She, 16] Sherwani, Shariq I., Haseeb A. Khan, Aishah Ekhzaimy, Afshan Masood and Meena K. Sakharkar, “Significance of HbA1c test in diagnosis and prognosis of diabetic patients”, Biomarker insights, Vol. 11, pp. 95-104, 2016. DOI: 10.4137/BMI.S38440.

[Shi, 17] Shickel, Benjamin, Patrick James Tighe, Azra Bihorac and Parisa Rashidi, “Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis”, IEEE Journal of Biomedical and Health Informatics, Vol. 22, No. 5, pp. 1589-1604, 2017.

[Shi, 18] Shi, Feifei, Qingjuan Li, Tao Zhu and Huansheng Ning, “A survey of data semantization in internet of things”, Sensors, Vol. 18, No. 1, pp. 1-20, 2018.

[Sho, 18] Shovon, Ahmedur Rahman, Shanto Roy, Tanusree Sharma and Md Whaiduzzaman, “A restful e-governance application framework for people identity verification in cloud”, International Conference on Cloud Computing, pp. 281-294. 2018. DOI: http://dx.doi.org/10.1007/978-3-319-94295-7_19.

[Sno, 12] Snoek, Jasper, Hugo Larochelle and Ryan P. Adams, “Practical Bayesian optimization of machine learning algorithms”, Advances in neural information processing systems, Vol. 25, pp. 1-9, 2012. DOI: https://doi.org/10.48550/arXiv.1206.2944.

[Sof, 21] Sofla, Maryam Sheikh, Mostafa Haghi Kashani, Ebrahim Mahdipour, and Reza Faghih Mirzaee, “Towards effective offloading mechanisms in fog computing”, Multimedia tools and applications, pp. 1-46, 2021. DOI: https://doi.org/10.1007/s11042-021-11423-9.

[Sor, 20] Sornalakshmi M., Balamurali S., Venkatesulu M., Navaneetha Krishnan M., Lakshmana Kumar Ramasamy, Seifedine Kadry, Gunasekaran Manogaran, Ching-Hsien Hsu and Bala Anand Muthu, "Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry", Neural Computing and Applications, pp. 1-14, 2020. DOI: https://doi.org/10.1007/s00521-020-04862-2.

[Sub, 19] Subramaniyaswamy V., Gunasekaran Manogaran R., Logesh V., Vijayakumar, Naveen Chilamkurti D., Malathi and Senthilselvan N., "An ontology-driven personalized food recommendation in IoT-based healthcare system", The Journal of Supercomputing, Vol. 75, No. 6, pp. 3184-3216, 2019.

[Sul, 14] Sultan and Nabil, "Making use of cloud computing for healthcare provision: Opportunities and challenges", International Journal of Information Management, Vol. 34, No. 2, pp. 177-184, 2014.

[Sun, 20] Suneetha, K. CH, R. Shanmuka Shalini, Vijaya Kumar Vadladi and Machunoori Mounica, "Disease prediction and diagnosis system in cloud based IoT: A review on deep learning techniques", Materials Today: Proceedings, pp. 1-5, 2020. DOI: https://doi.org/10.1016/j.matpr.2020.09.519.

[Tan, 17] Tang, Jie, Dawei Sun, Shaoshan Liu and Jean-Luc Gaudiot, "Enabling deep learning on IoT devices", Computer, Vol. 50, No. 10, pp. 92-96, 2017.

[The, 17] Thevenot, Jerome, Miguel Bordallo Lopez and Abdenour Hadid, "A survey on computer vision for assistive medical diagnosis from faces", IEEE journal of biomedical and health informatics, Vol. 22, No. 5, pp. 1497-1511, 2017.

[The, 21] Theis, Julian, William Galanter, Andrew Boyd and Houshang Darabi, "Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture", IEEE Journal of Biomedical and Health Informatics, Vol. 26, No. 1, pp. 388-399, 2021.

[Tka, 16] Tkac, Michal and Robert Verner, “Artificial neural networks in business: Two decades of research”, Applied Soft Computing, Vol. 38, pp. 788-804, 2016. DOI: https://doi.org/10.1016/j.asoc.2015.09.040.

[Ull, 17] Ullah, Farhan, Muhammad Asif Habib, Muhammad Farhan, Shehzad Khalid, Mehr Yahya Durrani and Sohail Jabbar, “Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare”, Sustainable cities and society, Vol. 34, pp. 90-96, 2017. DOI: https://doi.org/ 10.1016/j.scs.2017.06.010.

[Wal, 18] Walczak and Steven, “Artificial neural networks”, Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global, pp. 120-131, 2018.

[Wam, 13] Wamba, Samuel Fosso, Abhijith Anand and Lemuria Carter, “A literature review of RFID-enabled healthcare applications and issues”, International Journal of Information Management, Vol. 33, No. 5, pp. 875-891, 2013.

[Wan, 13] Wang, Wei, Suparna De, Gilbert Cassar and Klaus Moessner, “Knowledge representation in the internet of things: semantic modelling and its applications”, automatika, Vol. 54, No. 4, pp. 388-400, 2013.

[Wan, 16] Wang, Yasi, Hongxun Yao and Sicheng Zhao, “Auto-encoder based dimensionality reduction”, Neurocomputing, Vol.184, pp. 232-242, 2016. DOI: https://doi.org/10.1016/j.neucom.2015.08.104.

[Wan, 18] Wang, Feng, Haijun Liu and Jian Cheng, “Visualizing deep neural network by alternately image blurring and deblurring”, Neural Networks, Vol. 97, pp. 162-172, 2018. DOI: https://doi.org/10.1016/j.neunet. 2017.09.007.

[Was, 20] Washington P., Paskov K.M., Kalantarian H., Stockham N., Voss C., Kline A., Patnaik R., Chrisman B., Varma M., Tariq Q. and Dunlap K., “Feature selection and dimension reduction of social autism data”, Pacific Symposium on Biocomputing, pp. 707-718, 2020.

[WHO, 21] WHO, https://www.who.int/news-room/fact-sheets/detail/diabetes, 2021. Accessed on 13, December 2021.

[Xia, 17] Xia Y., Liu C., Li Y. and Liu N., “A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring”, Expert Syst. Appl, Vol. 78, pp. 225-241. DOI: https://doi.org/10.1016/ j.eswa.2017.02.017.

[Xia, 18] Xiao, Guohui, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati and Michael Zakharyaschev, “Ontology-based data access: A survey”, International Joint Conferences on Artificial Intelligence, pp. 5511-5519, 2018. DOI: https://doi.org/ 10.24963/ijcai.2018/777.

[Xie, 20] Xie, Huiqiang, Zhijin Qin, Geoffrey Ye Li and Biing-Hwang Juang, “Deep learning based semantic communications: an initial investigation”, Global Communications Conference, pp. 1-6, 2020. DOI: 10.1109/ GLOBECOM42002.2020.9322296.

[Yah, 19] Yahyaoui, Amani, Akhtar Jamil, Jawad Rasheed and Mirsat Yesiltepe, “A decision support system for diabetes prediction using machine learning and deep learning techniques”, International Informatics and Software Engineering Conference (UBMYK), pp. 1-4, 2019. DOI: 10.1109/ UBMYK48245.2019.8965556.

[Yan, 19] Yang, Yanqing, Kangfeng Zheng, Chunhua Wu and Yixian Yang, “Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network”, Sensors, Vol. 19, No. 11, pp. 1-20, 2019.

[Ye, 20] Ye, Jiancheng, Liang Yao, Jiahong Shen, Rethavathi Janarthanam and Yuan Luo, “Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes”, BMC Medical Informatics and Decision Making, Vol. 20, No. 11, pp. 1-7, 2020.

[Yil, 19] Yildirim, Ozal, Muhammed Talo, Betul Ay, Ulas Baran Baloglu, Galip Aydin and Rajendra Acharya U., “Automated detection of diabetic subject using pre-trained 2D-CNN models with frequency spectrum images extracted from heart rate signals”, Computers in biology and medicine, Vol.113, pp. 1-10, 2019. DOI: https://doi.org/10.1016/ j.compbiomed.2019.103387.

[You, 18] Young T., Hazarika D., Poria S. and Cambria E., “Recent trends in deep learning based natural language processing”, IEEE Computational intelligence magazine, Vol. 13, No. 3, pp. 55-75, 2018.

[You, 19] Yousif Ahmed, Safa and Hossana Twinomurinzi, “Exploring Semantic Interoperability in e-Government Interoperability Frameworks for intra-African collaboration: A Systematic Literature Review”, Hawaii International Conference on System Sciences, pp. 3182-3189, 2019. DOI: http://dx.doi.org/10.24251/HICSS.2019.384.

[Yu, 19] Yu, Eun-Seop, Jae-Min Cha, Taekyong Lee, Jinil Kim and Duhwan Mun, “Features recognition from piping and instrumentation diagrams in image format using a deep learning network”, Energies, Vol. 12, No. 23, pp. 1-19, 2019.

[Zak, 15] Zakir, Jasmine, Tom Seymour and Kristi Berg, “Big Data Analytics”, Issues in Information Systems, Vol. 16, No. 2, pp. 81-90, 2015.

[Zha, 18] Zhang, Libo, Tiejian Luo, Fei Zhang and Yanjun Wu, “A recommendation model based on deep neural network”, IEEE Access, Vol. 6, pp. 9454-9463, 2018. DOI: 10.1109/ACCESS.2018.2789866.

[Zho, 20] Zhou, Xiaokang, Wei Liang, Kevin I., Kai Wang, Hao Wang, Laurence T. Yang and Qun Jin, “Deep-learning-enhanced human activity recognition for Internet of healthcare things”, IEEE Internet of Things Journal, Vol. 7, No. 7, pp. 6428-6438, 2020.

[Zou, 18] Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang. “Predicting diabetes mellitus with machine learning techniques”, Frontiers in Genetic, Vol. 9, pp. 1-10, 2018. DOI: https://doi.org/ 10.3389/fgene.2018.00515.

Photocopies of
Papers Published in the
International Journals

Scopus Preview Author search Sources Create account Sign in

Source details

Annals of the Romanian Society for Cell Biology

Scopus coverage years: from 2009 to 2020
(coverage discontinued in Scopus)
Publisher: Annals of the Romanian Society for Cell Biology
ISSN: 1583-6258
Subject area: Biochemistry, Genetics and Molecular Biology: Physiology
Biochemistry, Genetics and Molecular Biology: Molecular Biology
Biochemistry, Genetics and Molecular Biology: Cell Biology
Source type: Journal

CiteScore 2020
0.6

SJR 2020
0.101

SNIP 2020
0.486

View all documents > Set document alert Save to source list Source Homepage

CiteScore CiteScore rank & trend Scopus content coverage

i Improved CiteScore methodology
CiteScore 2020 counts the citations received in 2017-2020 to articles, reviews, conference papers, book chapters and data papers published in 2017-2020, and divides this by the number of publications published in 2017-2020. Learn more >

CiteScore 2020

$$0.6 = \frac{\text{26 Citations 2017 - 2020}}{\text{44 Documents 2017 - 2020}}$$

Calculated on 05 May, 2021

CiteScore rank 2020

Category	Rank	Percentile
Biochemistry, Genetics and Molecular Biology — Physiology	#160/169	5th
Biochemistry, Genetics and Molecular Biology — Molecular Biology	#362/382	5th

View CiteScore methodology > CiteScore FAQ > Add CiteScore to your site

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

RULE EMBEDDED SEMANTIC ONTOLOGY BASED CLASSIFIERFOR IoTHEALTHCARE

S. Sathyapriya[1], Dr. L. Arockiam[2]

[1] Research Scholar, Department of Computer Science, St. Joseph's College, Tiruchirappalli,
[2] Associate Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli.
sathyapriya2822@gmail.com, larockiam@yahoo.co.in

Abstract

Internet of Things (IoT) is an emerging technology in all domains that generates large amounts of data at rapid pace. TheIoT devices are interconnected in a way to communicate and share data with each other. Knowledge mining from such large amounts of data is a difficult task. So commonly, data analytics modelsareused to extract knowledge. However, most data are not fully utilized because of their dynamic problems and difficulties in analyzing data collected fromdiverse resources. To overcome the above stated issues, semantic technologies are used to provide a common model to handle the data. In the field of healthcare, predicting the patient's disease accurately is one of the most important considerations. For this, semantic data is very useful to make accurate predictions quickly with minimal cost. In this paper, asemantic ontology based technique has beenproposed for IoT based healthcare domain. The proposed technique Rule Embedded Semantic Ontology Classifier (RESOC)is implementedin two steps,namely data collection and semantic enrichment. Data is collected through various sources and then the RESOC is developed in the semantic enrichment phase. Finally, theenriched semantic data enables theDeep neural Network (DNN)for disease classification. The resultsarecompared based on certain parameters such as precision, recall, F-score and accuracy. Hence, the semantically enriched ontology handles heterogeneity and improves classification accuracy.

Keywords: ***IoT, Semantic, Ontology, Prediction, Healthcare, Analytics***.

I. Introduction

Internet of Things (IoT) is a collection of sensing devices that can sense and communicate with each other through internet [1]. These sensing devices have the ability to share and receive information in a variety of applications and offer many services. IoT is one of the most popular internet technologies and has a wider population with real world [2]. Among various applications of IoT, Healthcare domain has driven more attention. IoT technologies provide efficient service to the healthcare domain like monitoring patient, diagnosing, giving treatment, takinga quickdecision,minimizing the cost andavoidingthe critical issue[3]. In general, IoT offers wearable devices for patientmonitoring and making decision about the current status of the patient. Nowadays, IoT medical sensors are used widely and it generates big volume of health-relateddata [4]. This data should be analyzed carefully because decisions are very important in the healthcare domain. But one of the major issues in this is data formats, because dataare collected from various sensors and in various sources. So, this will lead toheterogeneity, interoperability and scalability issues [5]. To avoid these issues, semantic web technologies are used toprovide aunique data model for data from various sensors. Normally, it converts the data to meaningful form and explores the structuresand relationshipsbetween the data.

Semantic refers to the meaning of datathat describes the single data in detail manner and it enables better communication by providing interoperability among devices [6]. Ontology plays a key role in semantics because it provides the explanation and characteristics of the data [7]. Another important technology is Resource Description Framework (RDF), it is used to provide a platform for semantic model and it enables the interoperability among various IoT devices [8]. Therefore, this work proposesa semantic based classification model for IoT based health care system. It annotates the datausing rule embedded ontology and returns data in the RDF triple format. Then the semantic information is extractedand given to the classifier that diagnosesthe disease accurately.

The remaining part of the paperis as follows, section II explores the existing research works relevant to the proposed work, section III explainsthe proposed RESOC technique for healthcare domain, Section IV discusses the results of the proposed architecture and Section V provides the results and limitations.

II. Related Works

The related works on IoT based semantic approaches are following,

Gergely Marcell et al., [9]reviewed the semantic sensor technologies in the internet of things. This work reviewed most widely or generally used ontologies in a summarized manner. He also explainedlayer wise semantic technologies forIoT systems. Finally, the authorconcluded that there was a need for more standardization so as to achieve flexibility, interoperability and quick results.

AhlemRhayem et al., [10] reviewed semantic web technologies in IoT environment. The proposed work reviewed the most relevant research in Semantic Web TechnologiesofIoT domain and summarized list of aspects & drawbacks. Finally, challenges and future opportunities were described.

NouraAlhakbaniet al.,[11]developed the event matching system (SMT) for semantic data in IoT context. The proposed algorithms matched events using a tree-based structure that supports systematic communication among critical applications. SMT was compared with existing work in terms of processing time, from which SMT achieved linear performance time. This system was not suitable for distributed environment and also parallel processing may be applied to improve the processing of event matching.

JoãoMoreiraet al., [12] proposed a SEMIoTICS model for early warning systems in internet of things. The proposedmodel provided semanticinteroperability forIoT systems and discussed some usecases. The model was validated by satisfying the requirements and overcame the challenges which were discussed.

M. Manonmani., [13] reviewed semantic annotation models for healthcare domain. This paper surveyed various data mining techniques which were used in healthcare domain as well as semantic annotation. The survey recommended solutions to overcome interoperability issues in healthcare domain by using semantic annotation models. Also, the steps which were involved in the semantic model creation using feature selection and classification algorithms were explained. This work has not reviewed many semantic annotation models.

SivadiBalakrishna et al., [14]proposed a work for data integration and data analysis using machine learning algorithms for IoT healthcare domain. Various semantic and machine learning techniques for data integration were reviewed. Moreover, future directions were discussed in the field of data integration from sensor in healthcare using semantic and machine learning approaches. The proposed approach for healthcare domain was not implemented using any tools.

T. Elsaleh [15], presented a lightweight IoT stream ontology for annotating streaming data. The model has been developed by following most recognized guidelines of semantic model and IoT environment. Thewell-known Semantic Sensor Networkontology for sensor descriptions was used in the developed light weight model. The annotated data were extracted in RDF Triple format and finally some use cases, tools, application were discussed. Scalability and quick processing were the essential parameters which have to be improved in this work.

Li Chen [16], developed an ontology-based model for diagnosing diabetes, monitoring and giving treatment to diabetes patients in a remote manner. The proposed ontology model solved the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

inconsistency problem by analyzing the patient information in detail. The performance of the proposed model was validated using Semantic Web Rule Language (SWRL) rules. Moreover, the experiment results proved that the model well predicted the diabetes disease and recommended prescriptions. The model was not suitable for critical situation because it took more time to process.

III. Methodology

This section proposesmethodology to track and monitor the patients' diseases and prescribe medicine.It includes two important phases, namely, User Module and Semantic Module. In the first phase, physician and patient communicate with each other with the help of IoT devices. The physicians can monitor patients remotely and prescribe medicines anywhere, anytime, without any restrictions. The semantic phase provides several facilities to handle the data from the heterogeneous devices. In semantic module, the dataare convertedinto RDF triple format. For this, rule embedded ontology is developed andis used to merge IoT data with healthcare domain information and find out the hidden relation among them. The semantic module handles the heterogeneity while dealing with various devices and directly interacts with the user module. It covers the semantics with the data by adding self-described information packages. The proposed RESOCtechniquewill handle heterogeneity and improve classification accuracy. The workingflow of theproposed techniqueis shown in Figure 1.

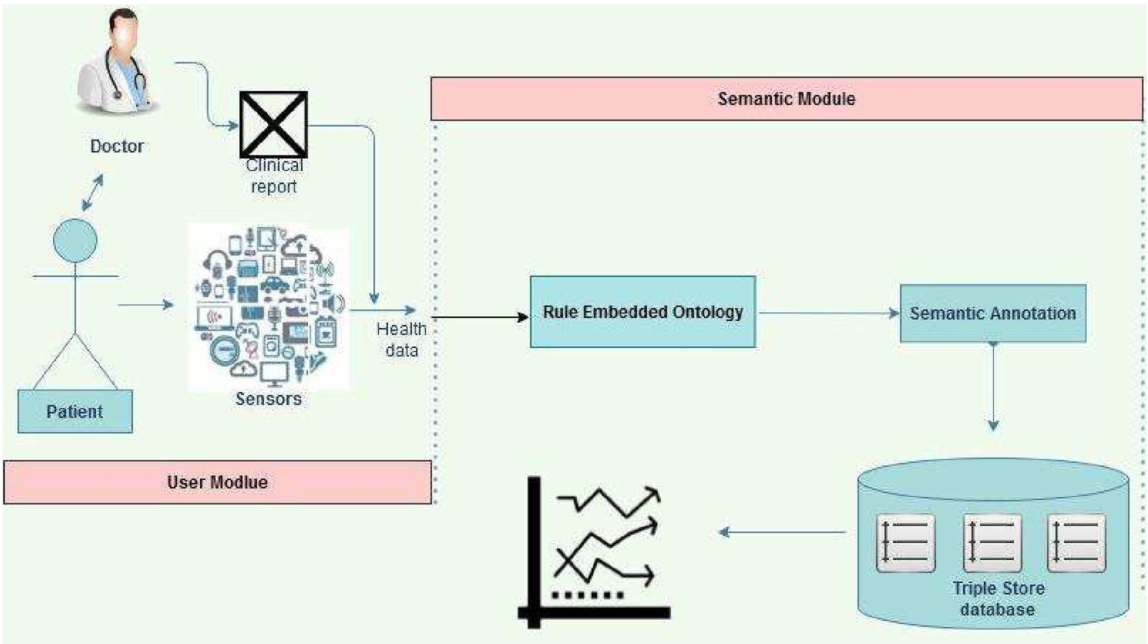


Figure 1: Proposed techniqueforIoT based healthcare domain

A. User Module

The primary role of this module is data collection, where the data arecollected through various methods likesensor mode, questionnaire mode and clinical report. Sensors play a vital role in data collection. There are various types of sensors used such as temperature sensor, heart rate sensor, air sensor, blood pressure sensors, and blood glucose sensor. The personal information collected through the questionnaire contains the details such as age, gender, food habits, heredity disease, height, weight, medicine intake, etc., some details from clinical reports like patient history are also collected. The collected details are sent to the semantic module, where the data is converted into meaningful and understandable format for quick analysis.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

B. Semantic module

Thismodule is developed for semantic enrichment of IoT data and semantically enriches data representations. Semantic technologies are applied on the collected data to form the metadata, which includes the environment, the sensor's interpretation, and the configuration to improve knowledge. Eachdata from the devices issent as a token to the semantic module where the rule embedded ontology processes the token and adds descriptions. As discussed earlier, ontology is the key concept of the semantic web that represents well defined knowledge and visualizes semantic descriptions. This phasedevelops rule embedded ontologythat drives semantic knowledge and represents described knowledge as a triple form using RDFgraph, namely subject, predicate and object. Subject denotes resources, object denotes values and predicate indicates the properties or features of the properties and reveals the relationship between subject and object. To develop the rule embedded ontology, basic terms and concepts have to be specified first and grouped into classes, subclasses, object properties and data properties. After that,a set of rules have to be embedded into the ontology. Some of the terminologies of proposed ontology are listed in table 1. Here, Actors are subjects, relations are predicates and concepts are objects.

Table 1: Terminologies for Ontology

Actors	Concepts	Relation
Patient, Doctor, Nurse, Physician, Staff, Admin, sensors, Manager	Fever, heart attack, treatment, blood pressure, high sugar, disease, eye disease, Temperature, Blood Pressure, Blood Glucose, air quality, etc.,	Has, affected by, treated by, has symptoms, has value, has tested, has risk, has side effects, etc.,

The ontology developed using Protégé 5.0 tool and the pictorial representation of it is shown in figure 2.

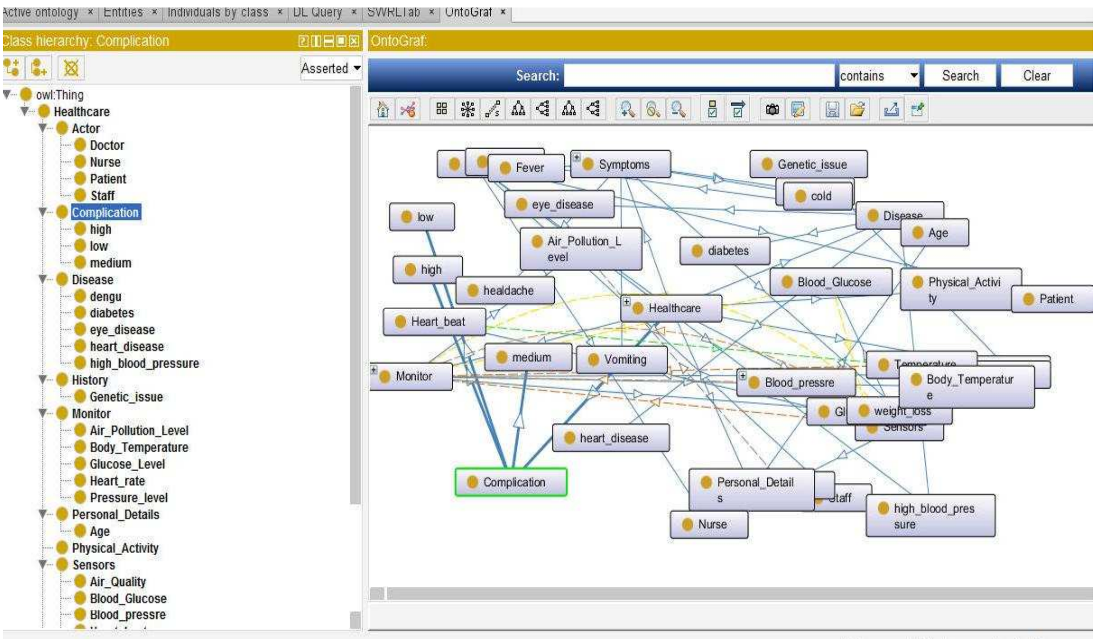


Figure 2:The rule embedded ontology for IoT Healthcare context

Many ontology based solutions are available for healthcare domain, but these are hard to implement with real time decision making. So, the rule embedded ontology is proposed whichperforms well in real time application. Here, the rules are developed using SWRL for finding hidden information from

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

individuals. Relation and concepts are considered as the primary sources of SWRL because these provide communication between two entities. More than 50 rules are generated based on the concepts of the ontology. Some of the rules are depicted in table 2. These rules play a vital role in semantic reasoning and disease classifications.

Table 2: Rules for ontology in SWRL format

Rules (Written in SWRL)	Descriptions
Patient(?, p)^has sensors(?p,temp)^Bodytemperature(?b)^diagnosis(temp , ?b)^has_value(?x,?v)^swrl:greaterThan(? V, 125)−>has_Complication (? P, High)	If the value of the temperature is higher than threshold value, then the patient has complication highest.
Patient(?, p) ^has_Symptoms (? P, weight loss) ^has_Symptoms (? P, frequent urination) −>has_Diagnosis (? P, Diabetes)	If the patienthas symptoms of weight loss and frequent urination then the patient has diabetes
Patient (? , p) has_booleanvalue (smoking, True) ^has_disease (? P, cough) −>has_Complication (? P, High)	If the patient has high complication due to the habit of smoking with cough
Patient (? , p)^has_sensors(?p,BP)^BloodPressure(?b)^diagnosis(pressureLevel, ?b)^has_value(?x,?v)^swrl:greaterThan(?v,150)−>has_Complication (?p, High)	If the value of blood pressure is higher than threshold value, and then the patient has complication highest.
Patient (? , p)^has_sensors(?p,BloodGlucose)^Gluoces(?b)^diagnosis(GlucoseLevel, ?b)^has_value(?x,?v)^swrl:greaterThan(?v,350)−>has_diagnosis (?p,Diabetes)	If the value of blood sugar is higher than threshold value, and then the patient has diabetes.

Then the Deep neural networks (DNN)classifier is applied to the semantically enriched data for classification.It permits to develop a model and defines its complex sequences in a simpler way.The semantically enriched data with a set of rules are given to the input layer of DNN and net input function X_{in} enables to combine all inputs with corresponding weights as defined in eqn(1). After that, Activation function $F(X_{in})$ is applied to process the inputs to provide the output. The following eqn(1) and (2) explains the construction of the classification model on semantically enriched data.

Let $cx_1.....cx_n$ be considered as inputs,

$$X_{in} = \sum_1^n cx_n w_n \text{ .... eqn(1)}$$

The above equ(1) is for computing inputs where, cx and w are inputs and its weights respectively, w provides the information about inputs.

Many Activation Functions are available in DNN. Here, linear activation function is used to process the inputs. It obtains inputs by multiplying w for each neuron and generates an output that is relative to the input vector. The output Y is formed as

$$Y(X) = F(X_{in}) \text{ .... eqn (2)}$$

The resulting output classifies the diseases. The performance of proposed RESOCtechnique is discussed in the section below.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

IV. Evaluation Results:

The proposed technique is evaluated based on parameters such asprecision, recall, F-score and accuracy.For comparison, the raw dataset and the dataset enriched with ontology are considered. To evaluate the performance, the following has to be identified;

TP: True Positive (TP) the quantity of accurately classified the positive labels
TN: True Negative (TN) the quantity of inaccurately classified the positive labels
FP: False Positive (FP) the quantity of accurately classified the negative labels
FN: False Negative (FN) the quantity of inaccurately classified the negative labels

a. **Precision**for finding the positive classified values from the total number of positive classes, defined as

$$\text{Pre} = \frac{TP}{TP+FP}$$

b. **Recall**also called as sensitivity, it calculates the positive class labels from the total number of class labels, that defined as

$$\text{Rec} = \frac{TP}{TP+FN}$$

c. **F-Score** is the harmonic mean of precision and recall

$$\text{F-S} = \frac{2(Pre*Rec)}{Pre+Rec}$$

d. **Accuracy**calculates the value of correct prediction on the dataset,

$$\text{Acc} = \frac{TP+TN}{TP+FP+TN+FN}$$

Figure 4 explains the comparative results of the classifier. The proposed RESOC technique enhances the classification accuracy of the DNN classifier. The results are computed based on the parameters that are discussed above. From which, the proposed RESOC technique provides 88% accuracy of the classifier.

The following figure 4is a graphical representation of the evaluation metrics.

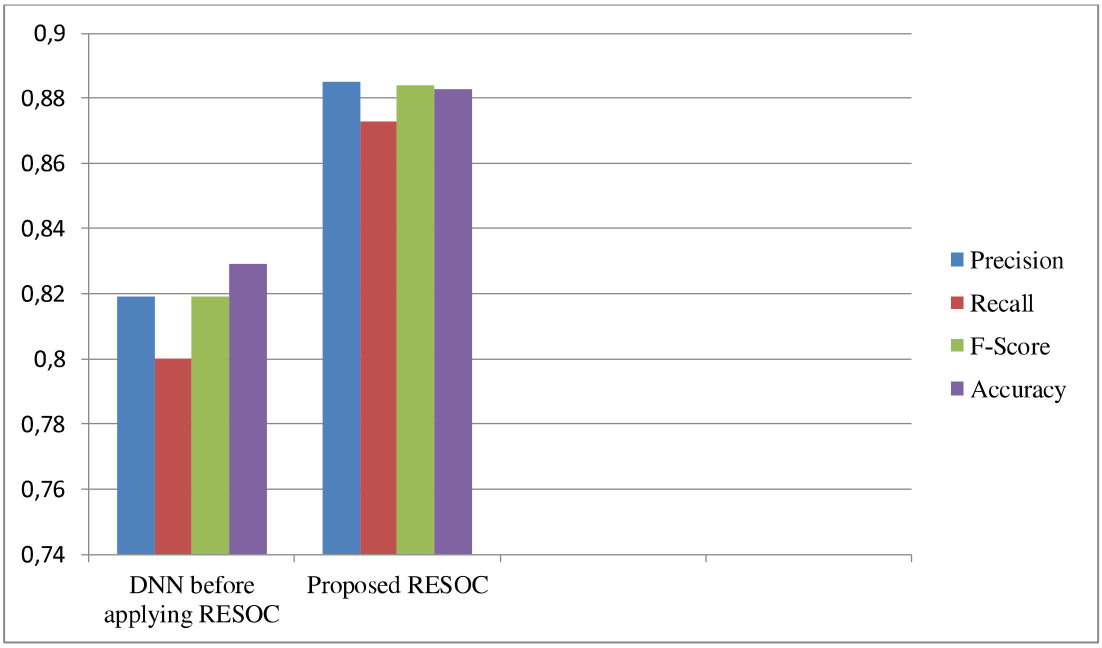


Figure 4: Graph representation of the evaluation results of RESOCtechnique

Conclusion

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 10224 - 10231
Received 05 March 2021; Accepted 01 April 2021

In this paper, RESOCtechniqueis proposed for handling IoT based healthcare domain data. The main goal of this work is to solve the heterogeneity problem and to improve classification accuracy. Thistechnique helps the physicians tomonitor the patients anytime, anywhere and at anyplace without any restrictions. It contains two modules, that are user module and semantic module. Data collection is done by various forms in the user module. In Semantic module, rule embedded ontology was developed to enrich the collected data and avoid heterogeneity.The enriched data was then classified using DNNand the performances were evaluated. Thus, the evaluation results prove that the proposed ontology improvesthe performance of the classifier but processing time is high. In future, the classifier will be enhanced to produce timely results.

References

[1]. Li, Shancang, Li Da Xu, and Shanshan Zhao, "The internet of things: a survey", *Information Systems Frontiers* , no. 2 ,Vol. 17, pp: 243-259, 2015.

[2]. Al-Fuqaha, Ala, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications" ,*IEEE communications surveys & tutorials*, no. 4 , Vol. 17, pp: 2347-2376,2015.

[3]. Karagiannis, Vasileios, PeriklisChatzimisios, Francisco Vazquez-Gallego, and Jesus Alonso-Zarate, "A survey on application layer protocols for the internet of things",*Transaction on IoT and Cloud computin,g,* no. 1, vol. 3, pp: 11-17, 2015.

[4]. Moustafa, Hassnaa, Eve M. Schooler, Gang Shen, and Sanjana Kamath, "Remote monitoring and medical devices control in eHealth", In *2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*,IEEE, pp. 1-8, 2016.

[5]. Mavrogiorgou, Argyro, AthanasiosKiourtis, KonstantinosPerakis, StamatiosPitsios, and DimosthenisKyriazis,"IoT in healthcare: Achieving interoperability of high-quality data acquired by IoT medical devices", *Sensors,* no. 9, vol. 19, pp: 1-24, 2019.

[6]. Burzlaff, Fabian, Nils Wilken, Christian Bartelt, and HeinerStuckenschmidt, "Semantic interoperability methods for smart service systems: A survey",*IEEE Transactions on Engineering Management, pp. 1*-15,2019.

[7]. Seok, Hyun-Seung, and Yong-Ju Lee, "Ontology-based IoT Context Information Modeling and Semantic-based IoT Mashup Services Implementation", *The Journal of the Korea institute of electronic communication sciences,* no. 4, vol.14, pp: 671-678, 2019.

[8]. Banane, Mouad, and AbdessamadBelangour, "A Survey on RDF Data Store Based on NoSQL Systems for the Semantic Web Applications," In *International Conference on Advanced Intelligent Systems for Sustainable Development*, Springer, Cham, pp. 444-451, 2018.

[9]. Honti, GergelyMarcell, and Janos Abonyi, "A review of semantic sensor technologies in internet of things architectures", *Complexity,* pp: 1-21, 2019.

[10]. Rhayem, Ahlem, Mohamed Ben Ahmed Mhiri, and FaiezGargouri, "Semantic Web Technologies for the Internet of Things: Systematic Literature Review", *Internet of Things* ,vol.11,pp:1-22, 2020.

[11]. Noura, Mahda, Amelie Gyrard, Sebastian Heil, and Martin Gaedke, "Automatic knowledge extraction to build semantic web of things applications", *IEEE Internet of Things Journal,* no. 5, vol.6, pp: 8447-8454.

[12]. Moreira, João LR, Luís Ferreira Pires, and Marten J. van Sinderen, "SEMIoTICS: Semantic Model-Driven Development for IoT Interoperability of Emergency Services", *ISCRAM,* pp:916-973, 2019.

[13]. Manonmani, M., and Sarojini Balakrishnan, "A Review of Semantic Annotation Models for Analysis of Healthcare Data Based on Data Mining Techniques", In *Emerging Research in Data Engineering Systems and Computer Communications*, Springer, Singapore, vol. 1054, pp. 231-238, 2020.

[14]. Balakrishna, Sivadi, M. Thirumaran, and Vijender Kumar Solanki, "IoT sensor data integration in healthcare using semantics and machine learning approaches", In *A Handbook of Internet of Things in Biomedical and Cyber Physical System*, Springer, Cham, pp. 275-300, 2020.

[15]. Elsaleh, Tarek, María Bermudez-Edo, ShirinEnshaeifar, Sahr Thomas Acton, RoonakRezvani, and P. Barnaghi, "IoT-Stream: A Lightweight Ontology for Internet of Things Data Streams", In*2019 Global IoT Summit (GIoTS)*, pp. 1-6, 2019.

[16]. L. Chen, D. Lu, M. Zhu, M. Muzammal, O. W. Samuel, G. Huang, W. Li, and H. Wu, "Omdp: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems," *International Journal of Distributed Sensor Networks*, no. 5,vol.15, p. 1–15, 2019

Scopus Preview   Author search   Sources   Create account   Sign in

Source details

Annals of the Romanian Society for Cell Biology

Scopus coverage years: from 2009 to 2020
(coverage discontinued in Scopus)

Publisher: Annals of the Romanian Society for Cell Biology

ISSN: 1583-6258

Subject area: Biochemistry, Genetics and Molecular Biology: Physiology
Biochemistry, Genetics and Molecular Biology: Molecular Biology
Biochemistry, Genetics and Molecular Biology: Cell Biology

Source type: Journal

View all documents >   Set document alert   Save to source list   Source Homepage

CiteScore 2020
0.6

SJR 2020
0.101

SNIP 2020
0.486

CiteScore   CiteScore rank & trend   Scopus content coverage

i Improved CiteScore methodology

CiteScore 2020 counts the citations received in 2017-2020 to articles, reviews, conference papers, book chapters and data papers published in 2017-2020, and divides this by the number of publications published in 2017-2020. Learn more >

CiteScore 2020

$$0.6 = \frac{\text{26 Citations 2017 - 2020}}{\text{44 Documents 2017 - 2020}}$$

Calculated on 05 May, 2021

CiteScore rank 2020

Category	Rank	Percentile
Biochemistry, Genetics and Molecular Biology — Physiology	#160/169	5th
Biochemistry, Genetics and Molecular Biology — Molecular Biology	#362/382	5th

View CiteScore methodology >   CiteScore FAQ >   Add CiteScore to your site

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

A Review Onsemantic Technology in Iot Healthcare

S. Sathyapriya[1], Dr. L. Arockiam[2],

[1] Research Scholar, Department of Computer Science, St. Joseph's College, Tiruchirappalli,

[2] Associate Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli.

Abstract

Internets of Things (IoT) devices are becoming more and more popular among the people. It enables services to applications by understanding a situation or context. IoT is widely used in many applications areas out of which the healthcare domain isthe most popular application. Here the patient related information is precious. The data emitted from the sensors is growing at a staggering pace. So, collecting and processing such huge amounts of data is the biggest task. The data from the different devices may vary by its type or nature, so it will lead to interoperability problem and heterogeneity issues. Semantics will help to solve these problems by providing various solutions. It combines domain and devices related information to the actual data.This paper has provided areview on semantic technologies-based health care system.Here,the background concepts and the importance of semantic technologies in IoT based applications are discussed. In future, a new semantic model with ontology will be proposed for processing the real time data in IoT-based health-care system.

Keywords: ***IoT, Semantic, Ontology, RDF, RDF Schema, OWL.***

Introduction

In the current era, IoT technology is causing a major revolution in real-world contexts [1]. Ittransforms the actual world substancesinto smart.IoT provides a platform for sensor/things/devices to enable free contact onelegantsituation and allows connecting with anyone, anytime, anywhere and any network [2]. IoT-based systems offers many applications like smart city, smart office, smart agriculture, smart retail, smart transportation, smart healthcare, smart water supply and smart energy etc., [4][5]. The data is collected through sensors and sent to the databases through embedded communication devices [6]. These communication devices exchange data by remote-controlled devicesand allow direct connection to improve quality of life. The continuity of the devices and objects are maintained bydiverse communication technologies such as ZigBee, Bluetooth, GSM and Wi-Fi, etc.,[7]. These IoT data should be handled properly, and then only the end users can take the correct decision or actions. The important phases to structure and compile the data transfer correctly are autonomous information gain, innovation, analysis and planning [8].Further, IoT has widespread coverage but

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

it is still in the infancy stage of research on various problems such as standardization, scalability, diversity, general service, domain specific service, speed of transition, data formats, integration and so on[9]. So, there is a need for new solutions to tacklesuch a huge data. Thispaper, introducesthe review on semantic technologies for IoT-based health care systems. SectionIexplains the basic concept of semantics, section II explores the related concepts on IoT semantics, section III explains the use of semantic approaches in healthcare domainand section IV discusses the ontologybased semantic approaches.

I. IoT semantics

The general definition of semantics is the meaning of words in any subject, language, programming languages, formal logic, and semiotics [10]. In IoT, it plays an important role in the knowledge system [11] that supports the enrichment of data, handles heterogeneity, interoperability and knowledge gaining from IoT data. In IoT, Semantics represents a single measurement data with metadata that includes, domain information, sensordepiction with its measurement to improve the understanding [12].Semantic web technology operates devices that understand data using ordered and machine understandableexplanations.The data description includes location, type, relation between the domain and context, data provider and different attributes of data [13]. The semantically annotated data aid to explain communication between IoT devices and other sources. So, multiple companies use semantic data to take effective business decisions and maintain results.

The most commonsemantic technologies are[14], Ontology, Resource Description Framework (RDF), Resource Description Framework Schema (RDFS), Web Ontology Language (OWL), **S**imple **P**rotocol **A**nd**R**DF **Q**uery **L**anguage (SPARQL), semantic annotations, and semantic reasoning. These technologiessupport integration and interoperability of sensor data.

Ontology [15] is a key concept in semantic technologies. It defines asa set of conceptual data or characteristics of data and showsthe relation between the data. In general, it describes the knowledge of a particular data by answering for what, where, when and how type of questions.Here, the relation may be between two things or devices or applications or services and a concept thatdefines a thing or some activity. The important components of ontology are classes, objects, attributes and association. In general, IoT ontology is categorized into three, namely, for domain ontology ,device ontologyand estimation ontology [16]. The Device ontologyrepresents the characteristic of sensors and actuators. The Domain ontology describes the real world physical concepts, measurements and associations with each other. The Estimation

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

ontology represents the services and its requirements. Ontology supports vocabularies to maintain meaningful communication between machines to machine. The ontology modeling provides a flexible framework to maintain knowledge, enrich sensor data and ensure interoperability.

The second important technology of semantics is RDF [17], It was launched by World Wide Web Consortium (W3C)in1999 to implant metadata, and outlined the available resources on the Internet. RDF supports interoperability directly on various IoT applications. RDF is sometimes called asa direct graph because it includes node set and predicted set. The syntactic structure of RDFincludes three views, namely, subject view, predicate view and object view.

The next technology is an RDF schema [18] which was developed for modelingvocabulary for data whichis supported by RDF. It isanextension of the RDF Glossary that outlinesassociated explanations, andestablishes meaningful connections among resources.

The fourth technology is Web Ontology Language (OWL) [19], introduced by W3C. OWL is a semantic markup language for exchanging the ontology on the network. OWL isconsidered as a family of languages that represents knowledge by writing ontology and enhances the glossary for RDF data. In addition, OWL includes three sub-languages that are OWL Full, OWL Description Logic (DL), and OWL Lite. Moreover, all these sub-languages are a syntactic extension of its antecedents.

The next technology, SPARQL [20],introduced by W3C and designed to perform RDF data,It is used to convert data into an RDF triple format and performs various operations like integration, querying and validation of the data.

Another important technology is Semanticannotation [21], which annotates data semantically to IoT properties.It is one of the most powerful mechanisms that enhance the semantic interoperability in IoT context. Adding semantic annotation toresource in IoT applications is a big task. Finally, semantic reasoning [22] is employed to predict a rational outcome from a collection of theories, facts and rules. This deals with rational properties and classes occurred on the RDF map installed within the ontology[23].

II. Role of SemanticsTechnologies in Internet of Things

Semantic technologiesessentials, challenges and its job in IoTdomain has been discussed by various authors.They are highlighted by following,

In [24]., the basic concept of IoT, semantic application and IoT-based semanticswere discussed. The concept of semantic technologies, reasoning and representation of data explained

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

the challenges that could occur in an IoT-based semantic application.In [25], semantic sensor technologieswere discussed for internet of things architectures.For which,most closely related paperswere surveyed and information extracted.The most widely or generally used ontologies were identified and summarized and concluded there was necessitate in achieveflexibility, interoperability and timely outcomes.In [26], the systematic literatureswerehavereviewedon the subject ofsemantic web technologies ofIoT environmentand most relevant works were discussed with regard to the challenges and future opportunities.Moreover, someresearchershave explained semantic based models to solve the problems which were faced in iot environment. Those are following,

João Moreiraet al., [27] proposed a SEMIoTICS model for early warning systems in the ~~i~~Internet of things. The proposed framework satisfied the semantic interoperability in IoT systems. Many use caseswere discussed and validated based on the performance. Some of the existingchallenges were solved by satisfying their requirements.

NouraAlhakbani et al., [28] developed the Event Matching System (SMT) for handling IoT semantic data. The proposed algorithms matched events using a tree-based structure that supports systematic communication among critical applications. SMT compared with existing work in terms of processing time, from which SMT achieved linear performance time. This system was not suitable for distributed environment and parallel processing.

MahdaNoura., [29] proposed a methodology to extract the existing ontology automatically. Word2vec and k-means machine-learning techniques were used in the ontology to analyze the concepts and propertiesstatistically. The designed methodology was employedinto three applications such as smart home, smart city and smart weather. Finally, it was evaluated, but there was no comparison.

III. Semantic Technologies in IoThealthcare domain

IoT plays a key role in all applications. Among them, there is a greater focus on the health sector. The use of sensors in the health field has increased and so has the evolution of health-related information.So, this section explains the uses of semantics in the healthcare domain.

In [30], thesemantic annotation models for healthcare domain were reviewed. This paper surveyed various data-mining techniques which were applied in thehealthcare sector as well as semantic annotation. The survey recommended solutions to overcome interoperability issues in thehealthcare domain by using semantic annotation models.   It alsoexplained the steps which

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

were involved in the semantic model creation using feature selection and classification algorithms. However, thisworkdid not clearly review semantic annotation models.In [31], two main contributionswereproposedusing semantic-based approaches for healthcare applications. The first work presented the semantic representation that includes various processes like data collection, pre-processing, extracting and semantic modeling. In the second, the IoT Medicare System was developed that integrated HealthIoT ontology. The proposed IoT Medicare system improved the decision making with data analyzing. The correctness of the proposed system was not analyzed.In [32], thesemantic interoperability model was developed for the IoT domainto avoid heterogeneity due to the various types of IoT devices. It was developed for healthcare domain where heterogeneous IoT devices wereused to monitor and communicate with patients. To share information between doctor and patient, a lightweight annotation model was created that semantically annotated the data. RDF converted the annotated data into triple format to make it more meaningful and SparQLqueryderivedthe knowledge. IoT data integration and fusion were not clearly defined and semantic reasoning needed tobe improved.

In[33], a semantic Medical IoT platform was proposed for IoT healthcare. The major goal of this work was to avoid heterogeneity, integrity and visualization issuesandthe proposed framework provided many solutions. Moreover, many services were offered, namely, localization, simplification and effective integration. In addition, contract-based security policies were introduced to ensure the confidentiality of patient in the healthcare domain. But, stability, usability and portability of the proposed system were not improved. In[34], the data integration and data analysis using machine learning algorithm and semantics technologies for IoT healthcare domain were discussed. Various semantic and machine-learning techniques for data integration were reviewed. Moreover, future directions were discussed in the field of data integration from medical sensors with semantic and machine learning techniques. The proposed approachfor healthcare domain wasnot implemented using tools. In [35], a similarity analyzer framework for solving the interoperability problems was introduced. These problems occurred when transforming the electronic record of the healthcare domain between institutions. For this problem, twoartificial intelligence algorithms were used namely, Word2Vec and Doc2Vec. These two algorithmswereutilized for figuring out the similar semantics of the numerical and textual in the dataset. The conflicts were removed and converted data into a common model. The parameter configuration and processing time of each algorithm were identified and accuracy

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

was calculated. The proposed framework was only suitable for text and numeric type of data in Electronic Health Record (EHR).

IV. Ontologies for IoT context

From among theexisting,there were several emerging ontologies created for the area of IoT, namely,SensorML , W3C Semantic Sensor Network(SSN) [36], [37], IoT.estontologies[37], IoT-Lite Ontology [38] andIoT-A information model [39].Inaddition, ontology-based semantic models were proposed to handle various issues in IoT. Some of them arethefollowing,

T. Elsaleh[40], presented a lightweight IoT stream ontology for annotating streaming data. The model has been developed by following most recognized guidelines of semantic model and IoT environment. The well known ontology SSN for sensor descriptions used to develop the light weight model. The annotated data were extracted in RDF Triple format and finally some usecases, tools, and applications in use were discussed. Scalability and short-time processing were the essential parameters were not focused in this work.

Li Chen [41], developed a model for diagnosing diabetes disease and monitoring patients remotely. The proposed ontology Model (ODMP) solved the inconsistency problem by analyzing the patient information in detail. The performance of ODMP model was validated using SWRL rules. Moreover, the experimental resultsproved that the model well predicted the diabetes disease and recommended prescriptions.However, the model wasnot suitable for critical situations because it took more time to process.

Ali et al.,[42] presented a recommendation system for IoT based healthcare sector. This work used health related data that from sensors. In addition, it continuously monitored the patient’s health andprescribedmedicines and food.This work only consideredmanagement and diet recommendation. Prediction and prevention of disease were not addressed.

Chen et al.,[43] suggested ontology for modelling the diabetic knowledge system. This knowledge system was prepared using health data. Based on the data, the disease was diagnosed and treatmentwas prescribed for diabetes. Moreover, the modelwas applied in disease prediction, diabetes diagnosis, and treatment recommendation. But, this work did not use IoT data.

Some of the IoT semantic-related works are discussed along with its drawbacks in the following Table 1.

Table 1: IoT semantic-related papers

Author	Year	Works Done	Drawbacks

I. Szilagyi [44]	2016	Discussed the importance of using ontologies in IoT.	There is no comparison or classification to this work.
G. Bajaj,[45]	2017	Overviewed diverse ontologies which consider in the IoT domain.	Semantic features were not considered to prepare ontologies.
M. Ganzha[46]	2017	Explained relevant studies in ontology and semantic interoperability.	IoT domain requirement were not discussed.
S. De , Y. Zhou[47]	2017	Reviewed IoT ontologies and categorized these approaches like general and domain levels. .	Reviewed only reusable ontology in WoT without considering other domains.
D. Andro ˇcec [48]	2018	Surveyedthe usage of SWT in IoT domain.	Semantic features were not comparedand WoT was not highlighted in this study.
K. Janowicz ,[49]	2018	Explained SSN ontology for IoT devices and actuators with SOSA.	The sevices of both sensor and actuator were not discussed.
K. Ryabinin[50]	2019	Developed IoT-based system. The work was done on the basis of SSN ontology, except that theywere newly created.	Reasoningand Semantic quality was not emphasized according to the assessment.

Conclusion

The main ambition of this paper was to emphasize thevarious semantic approaches for IoT context. This work describes the semantic approaches and their connection with healthcare data, the importance of ontology that contains annotation and interoperability issues among various devices that are utilized in themedical domain. Thus, the paper will be very useful to researchers who are interested in doing their research in the area of IoT Semantics. Also, aid to develop models that are useful for exchanging information between patient and doctors through the internet and take accurate decisionson time.

References

[1]. Li, Shancang, Li Da Xu, and Shanshan Zhao, "The internet of things: a survey", Information *Systems Frontiers*, no. 2 ,Vol. 17, pp: 243-259, 2015.

[2]. Al-Fuqaha, Ala, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications" ,*IEEE communications surveys & tutorials*, no. 4 , Vol. 17, pp: 2347-2376,2015.

[3]. Lin, Jie, Wei Yu, Nan Zhang, Xinyu Yang, Hanlin Zhang, and Wei Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications", *IEEE Internet of Things Journal* , no. 5, Vol.4, pp: 1125-1142, 2017.

[4]. Shah, Sajjad Hussain, and Ilyas Yaqoob, "A survey: Internet of Things (IOT) technologies, applications and challenges", *IEEE Smart Energy Grid Engineering (SEGE)*, pp. 381-385, 2016.

[5]. Ge, Mouzhi, Hind Bangui, and Barbora Buhnova, "Big data for internet of things: A survey" *Future generation computer systems,* pp: 601-614, 2018.

[6]. Karagiannis, Vasileios, Periklis Chatzimisios, Francisco Vazquez-Gallego, and Jesus Alonso-Zarate, "A survey on application layer protocols for the internet of things", Transaction *on IoT and Cloud computing*, no. 1, vol.3, pp: 11-17, 2015.

[7]. Siow, Eugene, Thanassis Tiropanis, and Wendy Hall, "Analytics for the internet of things: A survey", *ACM Computing Surveys (CSUR)*, no. 4, vol.51, pp: 1-36, 2018.

[8]. Marjani, Mohsen, Fariza Nasaruddin, Abdullah Gani, Ahmad Karim, Ibrahim Abaker Targio Hashem, Aisha Siddiqa, and Ibrar Yaqoob, "Big IoT data analytics: architecture, opportunities, and open research challenges", *IEEE Access* , vol. 5, pp: 5247-5261, 2017.

[9]. Kobusińska, Anna, Carson Leung, Ching-Hsien Hsu, S. Raghavendra, and Victor Chang, "Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing", *Future generation computer system*, vol.87, pp: 416-419, 2018.

[10]. Barnaghi, Payam, Wei Wang, Cory Henson, and Kerry Taylor, "Semantics for the Internet of Things: early progress and back to the future", International *Journal on Semantic Web and Information Systems (IJSWIS)*, no. 1, vol.8, pp: 1-21, 2012.

[11]. Wang, Wei, Suparna De, Gilbert Cassar, and Klaus Moessner, "Knowledge representation in the internet of things: semantic modelling and its applications", *automatika* , no. 4, vol. 54, pp: 388-400, 2013.

[12]. Bermudez-Edo, Maria, Tarek Elsaleh, Payam Barnaghi, and Kerry Taylor, "IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics", Personal *and Ubiquitous Computing*, no. 3, vol.27, pp: 475-487, 2017.

[13]. Palavalli, Amarnath, Durgaprasad Karri, and Swarnalatha Pasupuleti, "Semantic internet of things", In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pp. 91-95, 2016.

[14]. Dridi, Ahmed, Salma Sassi, and Sami Faiz, "A Smart IoT platform for personalized healthcare monitoring using semantic technologies", *IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1198-1203, 2017.

[15]. Kovács, László, and Edit Csizmás, "Lightweight ontology in IoT architecture", *IEEE International Conference on Future IoT Technologies (Future IoT)*, pp. 1-6, 2018.

[16]. S. Hachem, T. Teixeira, and V. Issarny, "Ontologies for the internet of things", *Proceedings of the 8th Middleware Doctoral Symposium*, ACM,NewYork, NY, USA, p. 3, 2011.

[17]. Sopek, Mirek, Przemyslaw Gradzki, Witold Kosowski, Dominik Kuziski, Rafa Trójczak, and Robert Trypuz, "GraphChain: a distributed database with explicit semantics and chained RDF graphs", In *Companion Proceedings of the The Web Conference*, pp. 1171-1178. 2018.

[18]. Hogan, Aidan, "RDF Schema and Semantics", *Springer*, Cham, pp. 111-183, 2020.

[19]. Anitha, V. , "A Survey Paper on Ontology Concepts in Semantic Web Technology and it's Applications", ieee, PP. 316 322, 2018.

[20]. Swathi, Gunnam, S. Mahaboob Hussain, Prathyusha Kanakam, and D. Suryanarayana, "SPARQL for Semantic Information Retrieval from RDF Knowledge Base", International Journal of Engineering Trends and Technology (IJETT), no.7, vol.41 , pp 351 354, 2016.

[21]. Al-Osta, Mahmud, Bali Ahmed, and Gherbi Abdelouahed, "A lightweight semantic web-based approach for data annotation on IoT gateways", *Procedia computer science*, vol.113, pp: 186-193,2017.

[22]. Maarala, Altti Ilari, Xiang Su, and Jukka Riekki, "Semantic reasoning for context-aware Internet of Things applications", *IEEE Internet of Things Journal*, no. 2, vol. 4, pp: 461-473, 2016.

[23]. Serrano, Martin, and Amelie Gyrard, "A Review of Tools for IoT Semantics and Data Streaming Analytics.", *Build. Blocks IoT Anal*, vol. 6, pp: 139-163, 2016.

[24]. Harlamova, Marina, Mārīte Kirikova, and Kurt Sandkuhl, "A survey on challenges of semantics application in the internet of things domain", *Applied Computer Systems*, no. 1, vol.21, pp: 13-21, 2017.

[25]. Honti, Gergely Marcell, and Janos Abonyi, "A review of semantic sensor technologies in internet of things architectures", *Complexity*, pp: 1-21, 2019.

[26]. Rhayem, Ahlem, Mohamed Ben Ahmed Mhiri, and Faiez Gargouri, "Semantic Web Technologies for the Internet of Things: Systematic Literature Review", *Internet of Things* ,vol.11,pp:1-22, 2020.

[27]. Moreira, João LR, Luís Ferreira Pires, and Marten J. van Sinderen, "SEMIoTICS: Semantic Model-Driven Development for IoT Interoperability of Emergency Services", *ISCRAM*, pp:916-973, 2019.

[28]. Alhakbani, Noura, Mohammad Mehedi Hassan, Mourad Ykhlef, and Giancarlo Fortino, "An efficient event matching system for semantic smart data in the Internet of Things (IoT) environment" *Future Generation Computer Systems*, vol.95, pp: 163-174, 2019.

[29]. Noura, Mahda, Amelie Gyrard, Sebastian Heil, and Martin Gaedke, "Automatic knowledge extraction to build semantic web of things applications", *IEEE Internet of Things Journal*, no. 5, vol.6, pp: 8447-8454.

[30]. Rhayem, Ahlem, Mohamed Ben Ahmed Mhiri, Mayssa Ben Salah, and Faiez Gargouri, "Ontology-based system for patient monitoring with connected objects", *Procedia computer science,* vol.112, pp: 683-692, 2017.

[31]. Manonmani, M., and Sarojini Balakrishnan, "A Review of Semantic Annotation Models for Analysis of Healthcare Data Based on Data Mining Techniques", In *Emerging Research in Data Engineering Systems and Computer Communications*, Springer, Singapore, vol. 1054, pp. 231-238, 2020.

[32]. Jabbar, Sohail, Farhan Ullah, Shehzad Khalid, Murad Khan, and Kijun Han, "Semantic interoperability in heterogeneous IoT infrastructure for healthcare", *Wireless Communications and Mobile Computing* ,pp.1-11, 2017.

[33]. Dridi, Ahmed, Salma Sassi, and Sami Faiz, "Towards a semantic medical internet of things", In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1421-1428, 2017.

[34]. Balakrishna, Sivadi, M. Thirumaran, and Vijender Kumar Solanki, "IoT sensor data integration in healthcare using semantics and machine learning approaches", In *A Handbook of Internet of Things in Biomedical and Cyber Physical System*, Springer, Cham, pp. 275-300, 2020.

[35]. Naveed, Arjmand, Yim Fun Hu, Tshiamo Sigwele, Ghulam Mohi-Ud-Din, Misfa Susanto, and Muhammad Ali, "Similarity Analyzer for Semantic Interoperability of Electronic Health Records Using Artificial Intelligence (AI)", *Journal of Engineering and Scientific Research,* no. 2,vol.1,pp: 53-59,2020.

[36]. Compton, Michael, Payam Barnaghi, Luis Bermudez, RaúL GarcíA-Castro, Oscar Corcho, Simon Cox, John Graybeal et al, "The SSN ontology of the W3C semantic sensor network incubator group", *Journal of Web Semantics, vol.*17 , pp: 25-32,2012.

[37]. Wang,W.; de, S.; Toenjes, R.; Reetz, E.;Moessner, K., " A Comprehensive Ontology for Knowledge Representation in the Internet of Things", In Proceedings of the 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, Liverpool, UK, pp. 1793–1798, 2012.

[38]. Bermudez-Edo,M.; Elsaleh, T.; Barnaghi, P.; Taylor," K. IoT-Lite: A lightweight semantic model for the Internet of Things", In Proceedings of the 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, France, pp. 90–97, 2016.

[39]. Botts, M.; Robin, A., " OpenGIS sensor model language (SensorML) implementation specification", OpenGIS Implementation Specification OGC **2007**, 7. Available online: http://www.opengeospatial.org/standards/sensorml (accessed on 1 October 2018).

[40]. Elsaleh, Tarek, María Bermudez-Edo, Shirin Enshaeifar, Sahr Thomas Acton, Roonak Rezvani, and P. Barnaghi, "IoT-Stream: A Lightweight Ontology for Internet of Things Data Streams", In*2019 Global IoT Summit (GIoTS)*, pp. 1-6, 2019.

[41]. Chen, Li, Dongxin Lu, Menghao Zhu, Muhammad Muzammal, Oluwarotimi Williams Samuel, Guixin Huang, Weinan Li, and Hongyan Wu, "OMDP: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems", *International Journal of Distributed Sensor Networks* , no. 5, vol.5, 2019.

[42]. Ali, S. R. Islam, D. Kwak, P. Khan, N. Ullah, S.-j. Yoo, and K. S.Kwak, "Type-2 fuzzy ontology–aided recommendation systems for iot– based healthcare," *Computer Communications*, vol. 119, pp. 138–155, 2018.

[43]. L. Chen, D. Lu, M. Zhu, M. Muzammal, O. W. Samuel, G. Huang, W. Li, and H. Wu, "Omdp: An ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems," *International Journal of Distributed Sensor Networks*, no. 5,vol.15, p. 1–15, 2019.

[44]. I. Szilagyi , P. Wira , "Ontologies and semantic web for the internet of things-a survey", *in: Industrial Electronics Society, IECON , IEEE*, pp. 6949–6954, 2016 .

[45]. G. Bajaj, R. Agarwal, P. Singh, N. Georgantas, V. Issarny, "A study of existing ontologies in the iot-domain", *arXiv*, pp: 1–24, 2017.

[46]. M. Ganzha , M. Paprzycki , W. Pawłowski , P. Szmeja , K. Wasielewska , "Semantic interoperability in the internet of things: an overview from the inter-iot perspective", *J. Netw. Comput. Appl,* vol.81, pp: 111–124, 2017.

[47]. S. De , Y. Zhou , K. Moessner , "Ontologies and Context Modeling for the Web of Things, in: Managing the Web of Things", *Elsevier,* pp. 3–36, 2017 .

[48]. D. Andro ˇcec , M. Novak , D. Oreški , "Using semantic web for internet of things interoperability: a systematic review", *Int. J. Semant. Web Inf. Syst. (IJSWIS)* , no. 4, vol. 14, pp:147–171, 2018.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 9810 - 9819
Received 05 March 2021; Accepted 01 April 2021.

[49]. Janowicz, Krzysztof, Armin Haller, Simon JD Cox, Danh Le Phuoc, and Maxime Lefrançois. "SOSA: A lightweight ontology for sensors, observations, samples, and actuators." *Journal of Web Semantics*, vol. 56 , pp: 1-10, 2019.

[50]. K. Ryabinin , S. Chuprina , K. Belousov , "Ontology-driven automation of IoT-based human-machine interfaces development",*International Conference on Computational Science, Springer*, pp. 110–124, 2019.

Source details

International Journal of Recent Technology and Engineering

Scopus coverage years: from 2018 to 2019

(coverage discontinued in Scopus)

Publisher: Blue Eyes Intelligence Engineering and Sciences Publication

E-ISSN: 2277-3878

Subject area: Business, Management and Accounting: Management of Technology and Innovation

Engineering: General Engineering

Source type: Journal

CiteScore 2018
0.0

SJR 2019
0.107

SNIP 2020
0.354

View all documents >   Set document alert   Save to source list   Source Homepage

CiteScore   CiteScore rank & trend   Scopus content coverage

i Improved CiteScore methodology

CiteScore 2018 counts the citations received in 2015-2018 to articles, reviews, conference papers, book chapters and data papers published in 2015-2018, and divides this by the number of publications published in 2015-2018. Learn more >

×

CiteScore 2018

$$0.0 = \frac{0 \text{ Citations } 2015 - 2018}{175 \text{ Documents } 2015 - 2018}$$

Calculated on 01 May, 2019

CiteScore rank 2018

Category	Rank	Percentile
Business, Management and Accounting └ Management of Technology and Innovation	#206/216	2nd
Engineering └ General Engineering	#276/283	1st

View CiteScore methodology >   CiteScore FAQ >   Add CiteScore to your site

International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-2, July 2019

Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India

Dr. L. Arockiam, S. Sathyapriya, V.A. Jane, A. Dalvin Vinoth Kumar

Abstract: ***Diabetes mellitus is a non-communicable disease, however it may lead to other health problems such as blood pressure, heart attack, vision problem, slow healing sores to patients with arthritis etc. Diabetes disease is caused due to lifestyle, food habits, and low level of fabrication of insulin and pedigree factors of individual. According to the study, there will be 552 million people around the world will be affected by diabetes at 2030. This paper estimates the total populations of type 2 diabetes patients in the central region (Cuddalore, Thanjavur, Perambalur, Tiruchirappalli, Ariyalur, Karur, Nagapattinam, Thiruvarur, Pudukottai, and Karaikal) of Tamil Nadu. Diabetes patients have been diagnosed with the help of various parameters such as blood pressure, body mass index, dietary history, physical activity and pollution level in the air. The Honeywell HPm series particle sensor is used to access the PM 2.5, PM 10 levels in the air. Considering the air quality as a parameter, there are lots of illnesses caused by air pollutants and also cause additional problems for people who are already suffering due to disease. This review work provides the knowledge about the prevalence of type-2diabetes and it will help people to take precautions about diabetes disease and its risk.***

Index Terms: Diabetes, Air Quality, Sensor, PM2.5, PM10.

I. INTRODUCTION

Diabetes mellitus is one type of non-communicable disease. The prevalence of diabetes is rapidly increasing all over the world at a tremendous rate [1].It occurs when the glucose level increases in the blood. Blood glucose is the main source which produces energy to human body. The high blood sugaris defined as a medical syndrome, which is also called as hyperglycemia, which is caused due to an inadequacy of insulin in the human body. The level of blood sugar is standardized by a hormone, which is done by the insulin generated by the pancreas. The pancreas is a very tiny organ which is placed between the stomach and liver that helps to digest the food. According to the report of World Health Organization (WHO)[2], the highest number of diabetes affected people are living in India. The total number of diabetes patients in the year 2016 is 7.8 million it will exceed 79.4 million by 2030. The International Diabetes Federation (IDF)[3] in the world has reported on diabetes that it has proved 425 million adults living with diabetes. According to the report of IDF, 5.2 % of Indian people are not aware that they are suffering from high blood sugar. In specific, the Madras diabetes research [4] foundation instructed that about 42 lakhs individuals are suffering from diabetes and 30 lakh people are in pre-diabetes.

A. *Types of diabetes disease:*

There are various ways to detect the presence of diabetes in the human body. There are three categories in diabetes mellitus. They are Type-1 diabetes, Type-2 diabetes and Gestational diabetes[5]. The early stage of diabetes is identified using the following factors such as long-lasting blood sugar, blood sugar fasting, diabetes history of genes, measuring waist and the ratio of height waist of individuals. In this paper type 2 diabetes is considered.

a. Type 2 Diabetes

Type 2 diabetes is called as non-insulin dependent diabetes[6]. In type 2 diabetes, pancreas produces sufficient insulin but the beta cells do not use it properly and that's why insulin resistance is caused. In such case, insulin tries to get glucose into the cell but it can't maintain instead of this the sugar level may increase in the blood. People may get affected by the type 2 diabetes at any age even in childhood. Type 2 [7] diabetes is caused by overweight and inactivity which leads to insulin deficiency. These types of diabetes can be controlled by weight management, regular exercise and nutrition. The symptoms of type2 diabetes are same as type 1diabetes except itching skin and the problem in vision. This type of diabetescan'tbe cured but can be controlled by medicine and injection which is given for diabetes, physical exercise, blood monitoring and glucose controlling.

B. *PubMed NCBI*

Over the past few years, awareness about diabetes is growing and the possibility also growing in this field. According to PubMed NCBI, referred as a journal for publishing MEDLINE papers, indexed by PubMed has computed diabetes related details which are surveyed from the year of 1983 and 2018 by using the keyword "Prediction and Diabetes". The surveyed results are shown in the form of graph, which is displayed in Fig1. The count for 2018 is extrapolated till June 27, 2018.

Revised Manuscript Received on July 20, 2019.

Dr.L.Arockiam, Associate Professor. Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

S.Sathyapriya, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

V.A.Jane, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

A. Dalvin Vinoth Kumar, Assistant Professor, REVA University, Bangalore.

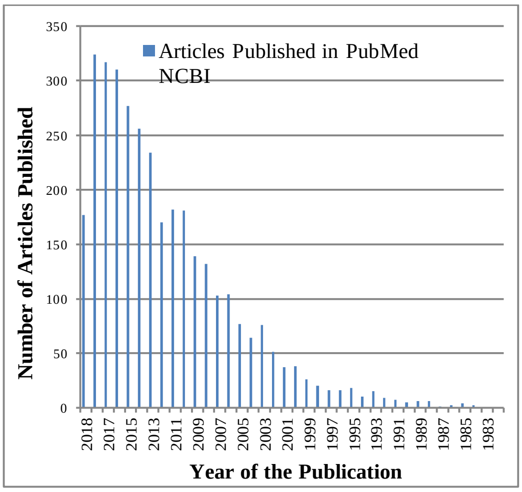


Fig 1: Number of publications index by Pub Med with keywords "Predictive and Diabetes." The 2018 count is extrapolated based on the number from June 27, 2018

II. REVIEW OF LITERATURE

Cheng lin et al [8]discussed about the classification and prediction in data mining by analyzing the information based on diabetes data. This paper partitioned the sets of data for classifying by using decision tree and data prediction was done through the linear regression, multiple regressions and non-linear regression whereas evaluated the classification accuracy. The process of classification and prediction of data mining also discussed about similarities and differences between them.

K.Lakshmi et al [9] proposed a System Architecture for diagnoses diabetes disease using clustering and classification algorithm such as decision tree and KNN. The proposed system has stored data into a server which was collected based on different diseases of patients. Here, they considered 11 attributes of diabetessuch as Age (years), Sex, Body mass index, Blood Pressure (mm Hg), Plasma Glucose Concentration (Glucose tolerance test) Triceps Skin fold 2-Hour serum insulin Diabetes Pedigree function, Cholesterol Level, Weight (kg) and Class variable (0 or 1) to predict the diabetes.The proposed method consists of some basic components such as admin, user (doctor, patient, physician etc), server, database, application, and data mining techniques. In the first step of the proposed system, KNN and Decision tree were applied for training the dataset after receiving the request from the user, which are like a supervised classification model.Admin received the inputs from requestor. In the final step DM approach was used to predict the result and send back to the user. Time and cost are reduced to diagnoses in this approach.

Dr.Prof.Neeraj et al [10] described the J48 algorithm for predicting recurrence of cancer-based data set to breast cancer. Recurrent cancer can be analyzed in three ways and they are: cancer comes back after treatment or it is in the same place, where it started first whether in any portion of the body. Hereafter J48 algorithm was used on the data set of breast cancer and implemented by WEKA tool and generated the decision tree by using 10 fold cross-validation method to predict the recurrent event due to its attributes such as tumor size, the degree of malignancy, age, node-caps, menopause etc.UCI machine learning repository provided the data set for predicting recurrence cancer of undergone treatment to patients. A result of experiment was tabulated and the decision tree was shown in the figure. Furthermore, results were concluded accurately and specific range value was used to find out the changes of recurrence cancer.

Manal Abdullah et al [11] proposed a method for finding five types of anemia is one of the hematological diseases and predicted what type of anemia hold by patient using classification algorithms. This paper proposed an algorithm for classification with the help of complete blood count test. The data sets were collected from patients and were filtered. Multiple experiments were conducted using various algorithms namely naive Bayes, neural network, J48 decision tree, and SVM. Compared with other algorithms J48 decision tree provided the best potential classification of anemia types. J48 decision tree algorithm provided better performance with accuracy, recall, true positive rate, false positive rate, precision and F-measure and it was proved by weka experiment. The tested results were tabulated in percentage (like 20%, 40%,60%). The anemia types can be detected with the help of given algorithms but this paper concentrated only on five types of anemia for finding accuracy and prediction of preferred results.

Himansu Das et al [12] focused on Diabetes Mellitus Disease. They used two data mining technique such as J48 and Navie Bayesian for predicting diabetes. The proposed technique was quicker and efficient for diagnosis the disease. The dataset was collected from medical college hospital by providing set of questions that about particular patient name, age, sex, blood, sugar level, and plasma glucose and as well as online repository. After thatthe data cleaning was performed to remove the unnecessary data and was stored in the warehouse. The proposed method predicted whether the patient has diabetes or not, by classification technique. The two classification techniques were implemented through WEKA software and the experimental results were tabulated. Navie Bayes better than J48 and also the outcome was proved by its productivity.

N.Vijayalakshmi and T.Jenifer [13] worked on data mining and statistical analysis for identifying diabetes disease. The data source contained pertaining diabetes which has taken from nursing home research center. The collected data divided as diabetic patients and non-diabetic patients. WEKA tool was used for analyzing the most important factors causing diabetics and also used to perform statistical analysis method on every single attribute. Tow classification techniques such as J48 pruned tree technique and the Random tree provided the validation result and the detailed accuracy on datasets by class. Hence this paper proved J48 pruned tree is a better technique compared with other classifying techniques and the accuracy of the predicted result was 81%.

III. SURVEY AREA

Tamil Nadu is one of the states in India. Based on the direction of the districts located, it is divided into 4 Regions namely central region, western region, southern region and Chennai city region. Each region has at least more than 4 districts. The central region has 10 districts such as Cuddalore, Thanjavur, Perambalur, Tiruchirappalli, Ariyalur, Karur, Nagapattinam, Thiruvarur, Pudukottai, and Karaikal. The western region has 6 districts which are Coimbatore, Erode, Nammakal, Salem, Dharmapuri and the Nilgiris. The southern region has 9 districts that are Dindigul, Madurai, Theni, Sivaganga, Virudunagar, Ramanathapuram, Tirunelveli, Thoothukudi and Kanyakumari. Finally, Chennai, Thiruvalluvar, Kancheepuram, Vellore, Tiruvannamalai, and Puducherrydistricts have come under the Chennai city region.

A. *Central region*

According to the census report at 2011, the Central region's total population is 12,212,084 where the men and women are in the frames of 7,031,520 and 7,194,867 . The total taluk in the central region of all districts are 66 whereas total revenue villages and panchayat villages are 4638 and 3154 respectively. From the report, the total number of literate people in that region is 7,369,787. Men and women in this category are 3,982,437 and 3,432,656. The total number of children (age between 0-6) in this region is 1,042,373, from this total number of male children and female children are 3,982,437 and 3,432,656.

IV. MATERIALS AND METHODS

All the study samples were randomly collected from states in the central region of Tamilnadu. The total study population is 10115 among them 5566 were male and 4549 femlae which is 55.1% and 44.9% respectively. The population was screened for blood pressure (diastolic and systolic) and blood sugar along with their screening data, the body mass index (BMI), dietary history, physical activity, pattern and Pm2.5 (pm & Pm10). The population screened for diabetics by random Blood Sugar Meter(RBS). The Blood pressure is screened using Arm Bp digital monitor. The dietary history, physical activity are assessed by a set of stored questions. The air pollutants (Pm2.5 & Pm10) are assured using Honeywell HPm series particle sensor. The number of total study population for male and female percentage has separated based on their age wise and listed in the table 1.

Table 1: Age and sex wise distribution of the study population

Age	No. Male Population (%)	No. Female Population (%)	Total Population (%)
< 30 years	1422 (72.9)	526 (27.1)	1948 (100)
30-35 years	1658 (60.7)	1071 (39.3)	2729 (100)
36-40 years	1427 (69.3)	632 (30.7)	2059 (100)
41-50 years	612 (36.25)	1076 (63.75)	1688 (100)
51-60 years	376 (23.7)	1213 (76.3)	1589 (100)
>60 years	71 (69.7)	31 (30.3)	102 (100)

V. RESULTS AND DISCUSSION

According to the report of total study population, people have separated based on their age and sex. From this, the total number of male and female has displayed in Fig2 in the form of graph. The age of both gender classified as, Below 30, 30 to 40, 41 to 50, 51to 60 andAbove 60.

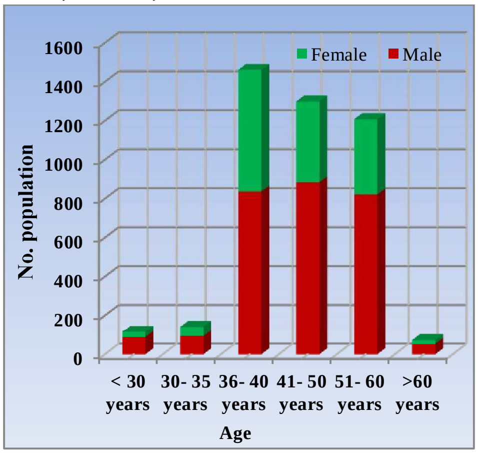


Fig 2: Distribution of Age and Sex

A. *Diabetic and Age*

Among the major factors of diabetes, age is considered like one kind of major factor. The total number of diabetes patients derived from total study population has given in the graph with its percentage. Fig3 represents the above mentioned details as a graph.

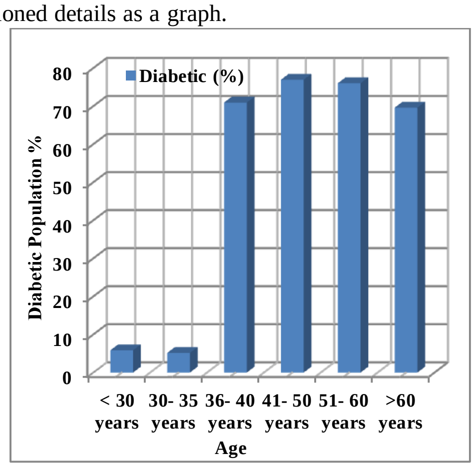


Fig 3: Diabetic and Age

B. *Education and Diabetic*

Diabetes awareness between literate and illiterate were surveyed. Totally 36.50 % percentage of illiterate people has lived in Tamilnadu, 41% percentage of people completed their schooling,22.50% percentage completed graduation. The comparison is between these categories of people represented in the form of graph in Fig4.

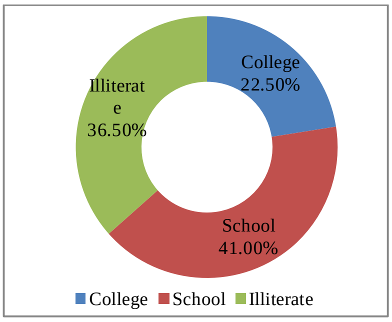


Fig 4: Education and Diabetic

C. *Air Pollution and Diabetic*

When the pancreas function decreases, the function of insulin is reduced. Diabetes occurs when the pancreas does not produce sufficient insulin. Today Air pollution is increasing throughout the world, and the air is most often polluted by the urban area so air pollution may affect the pancreas as well as the **livelihood may be affected to diabetic patients**. Here using Honeywell HPm series particle sensor, the air pollution(Pm 2.5 and Pm10) detail was collected and displayed. An average of air pollutants level is given as a graph in fig 5.

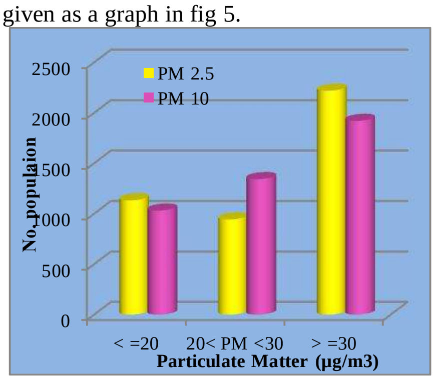


Fig 5: Air Pollution and Diabetic

D. *Diabetic Control measure*

According to this study, diabetes people have followed insulin or treatment taken from required government hospital or have followed any diet to control their diabetes or awareness about HCA1C test and carbohydrate count. In order to the study of total population has described and the number of population based on diabetic control measures which is represented in fig 6 as a graph.

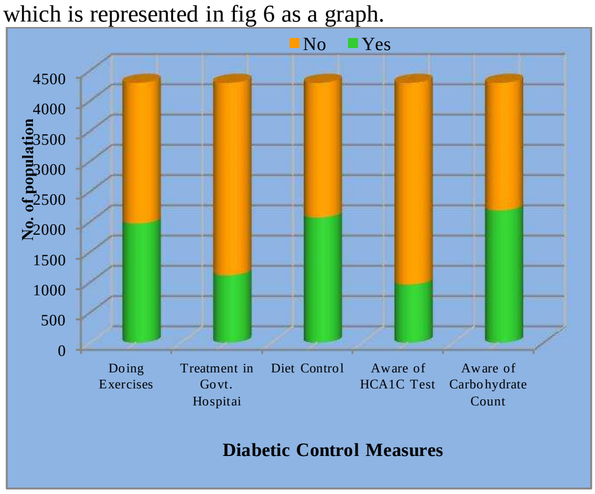


Fig 6: Diabetic Control measure

VI. CONCLUSION

Diabetes Mellitus is a chronic disease that can affect human life. Massive data was collected from census report at 2011, questionnaire and IoT devices. Tamilnadu has been separated into four regions with respect to the location. This study concentrated mainly on the central region of Tamilnadu and total population was surveyed in that region. From this, total number of male and female population was also reviewed. The number of people living with diabetes in the central region was calculated by using various parameters. The results of experiments exhibited the number of diagnosis made for diabetic patient and were computed individually on the basis of their age, education, physical activity, dietary history, and air pollution. This paper will help to spread the awareness about diabetes among people. In future, these experiments may be conducted all over Tamilnadu and it may improve the accuracy level with the help of various parameters

ACKNOWLEDGMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).Authors thankful to Dr. Ravi MD, CEO,MHealth,Trichy Tamil Nadu India and MrPeriyasamy, Meyer pharmaceuticals for helping us in data collection and also thankful to all the patients who have participated in this survey.

REFERENCES

1. P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," Int. Res. J. of Eng. and Tech. IRJET, Vol. 02, pp. 1039-1043, June-2015.
2. NDTV Food Desk, Updated: November 14, 2017 13:06 IST, available at: https://www.ndtv.com/food/world-diabetes-day-2017-number-of-diabetics-to-double-in-india-by-2023-1775180.
3. World Health Organization. "Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus", 1999, pp. 17-21.
4. American Diabetes Association. "Diagnosis and classification of diabetes mellitus", Vol.37, Issue.1, 2014, pp : 81-90.
5. Olson, Brooke. "Applying medical anthropology:Developing diabetes education and prevention programs in American Indian cultures", American Indian Culture and Research Journal, Vol.23, Issue.3, 1999, pp: 185-203.
6. Alberti, G., Zimmet, P., Shaw, J., Bloomgarden, Z., Kaufman, F. Silink, M. "Type 2 diabetes in the young: the evolving epidemic Diabetes care", Vol.27, Issue.7,2004, pp: 1798-1811.
7. Recognising Type-2 Diabetes', Feb 2016 Available: https://www.healthline.com/ health/type-2 -diabetes /recognizing-symptoms, [Accesed : 15-jan-2018].
8. Lin, C., & Yan, F. "The study on classification and prediction for data mining" In Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference ,2015,pp. 1305-1309.
9. Dr Prof. NeerajBhargava, N., Sharma, S., Purohit, R., &Rathore, P. S. , "Prediction of recurrence cancer using J48 algorithm" Proceedings of the 2nd International Conference on Communication and Electronics Systems, 2017,pp:386-390.
10. K. Lakshmi,D.Iyajaz Ahmed & G. Siva Kumar, " A Smart Clinical Decision Support System to Predict diabetes Disease Using Classification Techniques" IJSRSET,vol.4,Issue.1,2018,pp: 1520-1522.
11. Abdullah, M., & Al-Asmari, S.," Anemia types prediction based on data mining classification algorithms", Communication, Management and Information Technology–Sampaio de Alencar (Ed.),2017,pp:615-621

12. Himansu Das, BighnarajNaik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", Springer Nature Singapore Pte Ltd,2018,pp:539-549.
13. Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",International Journal of Computer Science and Mobile Computing, Vol.6, Issue.7, 2017, pp:166 – 172.

AUTHORS PROFILE

First Author Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.

Second Author S.Sathyapriya is doing her Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. Her research area is IoT Data Analytics.

Third Author V. A. Jane is doing his Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. His research area is IoT Data Analytics.

Fourth Author Dr. A. DalvinVinoth Kumar is working as Assistant Professor in the Department of Computer Science, Kristu Jayanthi College Bengaluru, Karnataka, India. His research interests are: MANET, Routing and IoT

Scopus Preview Author search Sources Create account Sign in

Source details

International Journal of Recent Technology and Engineering

Scopus coverage years: from 2018 to 2019

(coverage discontinued in Scopus)

Publisher: Blue Eyes Intelligence Engineering and Sciences Publication

E-ISSN: 2277-3878

Subject area: Business, Management and Accounting: Management of Technology and Innovation

Engineering: General Engineering

Source type: Journal

View all documents > Set document alert Save to source list Source Homepage

CiteScore 2018
0.0

SJR 2019
0.107

SNIP 2020
0.354

CiteScore CiteScore rank & trend Scopus content coverage

i Improved CiteScore methodology

CiteScore 2018 counts the citations received in 2015-2018 to articles, reviews, conference papers, book chapters and data papers published in 2015-2018, and divides this by the number of publications published in 2015-2018. Learn more >

CiteScore 2018

$$0.0 = \frac{\text{0 Citations 2015 - 2018}}{\text{175 Documents 2015 - 2018}}$$

Calculated on 01 May, 2019

CiteScore rank 2018

Category	Rank	Percentile
Business, Management and Accounting — Management of Technology and Innovation	#206/216	2nd
Engineering — General Engineering	#276/283	1st

View CiteScore methodology > CiteScore FAQ > Add CiteScore to your site

Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning

L. Arockiam, S. Sathyapriya, V.A. Jane, A. Dalvin Vinoth Kumar

Abstract: Machine learning is a part of AI which develops algorithms to learn patterns and make decision form the massive data. Recently, Machine learning has been used to resolving various critical medical problems. Diabetes is one of the dangerous disease, which can lead to more complicated, including deaths if not timely treated. The study is designed for providing the prevalence of Diabetes Mellitus in Tiruchirappalli district using machine learning algorithms and it was detected that the polluted air causes diabetes disease and also increases the risk of that disease. This proposed work helps the people in preventing diabetes disease using various diabetic attributes with an aim to enhance the quality of healthcare and lessen the diagnoses cost of the disease. In future, the work done may be extended by considering many other attributes and by implementing it through various algorithms to improve the prediction accuracy of diabetes mellitus.

Index Terms: Diabetes Mellitus, Machine Learning, Prediction, WEKA .

I. INTRODUCTION

Machine Learning plays an efficient role in medical especially diabetes research. Diabetes is a widely spreading disease in this modern society due to exercise gap, increased obesity rates, food habits and environment pollutants etc. Research on diabetes plays an important role in the field of medicine, and the number of daily data in this field is high. Continuous measurements are best suited for implementation of these data using data mining methods and can be handled immediately and these methods differ from other traditional methods and also one of the best ways in diabetes research when handle massive amounts of data related to diabetes. The main difference between them is more complicated than statistical approaches. Every day vast amount of data are stored in the various domains like finance, banking, hospital, etc. and rapidly increasing day by day. Such a Database may contain potential data that can be useful for decision making. Extraction of this valuable information manually from large volume of data is extremely difficult task. From the rapidly growing data, it is very hard to find useful knowledge without using ML techniques. Discovered knowledge can be useful in making prominent decisions. Data mining is widely used in fields such as business, medicine, science, engineering and so on [1-5].

Revised Manuscript Received on July 20, 2019.

Dr.L.Arockiam, Associate Professor. Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

S.Sathyapriya, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

V.A.Jane, Ph.D Scholar, Department of Computer Science, St. Joseph's College(Autonomous), Trichy-2.

A. Dalvin Vinoth Kumar, Assistant Professor, REVA University, Bangalore.

II. RELATED WORKS

Himansu Das et al., [6] proposed a framework for predicting diabetes mellitus. Diabetes Mellitus was predicted by classification algorithms such as j48, Naïve Bayes and these two were implemented using the weka tool. Questionnaire based data collection was done and data cleaning was performed to remove the unwanted data. The diabetes mellitus had been diagnosed by using j48 and Naïve Bayes. The final stage in the proposed framework generated the report of diabetes.

N.Vijayalakshmi and T.Jenifer [7] analysed risk factors of diabetes through data mining and statistical analysis techniques. The experiment for diabetes prediction was done by using classification algorithms, clustering, and subset of evaluation, association rule mining and statistics analysis. J48 provided better accuracy of 81% to the given dataset than the other techniques.

C.Kalaiselvi and G.M .Nasiria [8] predicted whether people with diabetes may have cancer and heart disease. Diabetes dataset was classified by using ANFIS and AGKNN algorithm and gained good accuracy level. The performance of algorithms was evaluated by using performance metrics. The proposed method reduces the complexity than the exiting methods.

Swaroopa shastri et al., [9] proposed a system to predict whether type 2 diabetes influences kidney disease. Here by the data mining algorithms were utilized. The proposed system generated the report of a patient, it assisted doctors, and also suggested precautions to the patient from kidney disease.

Huwan- chang et al., [10] developed a model for predicting postprandial blood glucose to undiagnosed diabetes cases in a cohort study. For this purpose, there were five data mining algorithms that were utilized and compared each other in this work. The data set used in this model was collected from Landseed Hospital in northern Taiwan over the period of 2006 to 2013 and also evaluated the performances of the data mining algorithms. The overall result of the proposed model provided the accurate reasoning and prediction; it could be useful to assist doctors to improve the skill of diagnosis and prognosis diseases.

Aiswarya Iyer et al., [11] utilized Decision Tree and Naïve Bayes algorithms for predicting diabetes in pregnant women. Training and test data was separated by 10 fold cross validation technique and J48 algorithm was employed on the Pima Indians Diabetes Database of "National Institute of Diabetes and Digestive and Kidney Diseases" using WEKA. The proposed work concluded that both algorithms were efficient for the diagnosis of diabetes and Naïve Bayes technique gave the result with least error rate.

A.A. Aljumah et al., [12] recommended a model based on regression technique for diabetes treatment. The proposed model predicted the diabetes disease by Oracle Data Miner tool and results were employed for experimental analysis on collected Datasets by support vector machine algorithm (SVM).

Mohammed et al., [13] presented a survey on application using Map Reduce programming framework which was discussed in early work and discussed Hadoop implementation in clinical big data related to healthcare fields.

N.M. Saravana Kumar et al., [14] proposed a Predictive Analysis System Architecture with various stages of data mining. Prediction approach carried out on Hadoop / Map Reduce environment. Predictive Pattern matching system was used to compare the threshold value analyzed with the estimated value after the analyzed reports were presented by the system.

III. METHODOLOGY

The proposed Model plays a significant role in predicting diabetic patients and produces the prevalence report of diabetes.

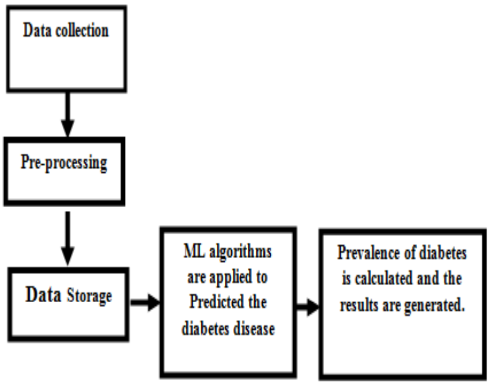


Fig 1: Work flow of proposed methodology

The work flow for diabetic prediction is shown in fig 1. In the initial step, the data collection is performed and it done through various ways such as questionnaire based data collection, sensor based data and some data from clinical report. Cloud storage is used where the electronic records are stored securely and cloud computing can be utilized for: data processing, data analysis and predictive analysis. These are carried out by statistical tools and data mining techniques. The predicative analytic stage sends the report of diabetes prevalence in Tiruchirappalli.

1. Data collection: It is one of the most initial steps in the proposed model and plays a major in data related research. In this paper there were following three types of data format collected from sensors, clinical and questionnaire.

2. Questionnaire: The data collected through questionnaire is called as the primary data. There were two types of data that were collected namely medical data and personal details. The questionnaire was prepared and given to various people who are living in Tiruchirappalli district. The question was developed using Google Form with 22 questions based on various factors such like gender, habits which spoils their health like smoking and alcohol drinking, food habit, BMI, medication taken by individual, blood pressure, family history , sleeping time, normal health problem , work type , educational background, environment pollutants and physical activity. Some of the questions were in yes/no format and some were in answer format. The model of the questionnaire sheet is given below in fig 2.

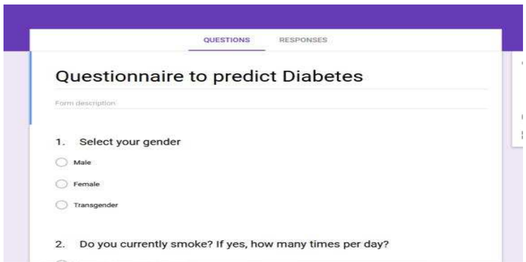


Fig 2: Questionnaire model based data collection

3. Sensor Data: Some data were collected by using sensor and also by using medical devices. In this thesis, Honeywell HPm Particle Sensor is used to find out the PM 2.5 and PM10 in the air and it is shown in fig 3. PM means particulate matter it used to find out the particles level in the air. PM 2.5 means particles with a size below 2.5 microns and PM10 includes particles with 10 microns and below. PM 2.5 is very serious than PM10 because PM2.5 contain very small particles it can travel to our lungs deeply and then causes more harmful effects. Further, it can lead to diabetes. In this paper particle matter is considered as a factor to predict the diabetes disease because air is an important factor for the people to survive in the world.

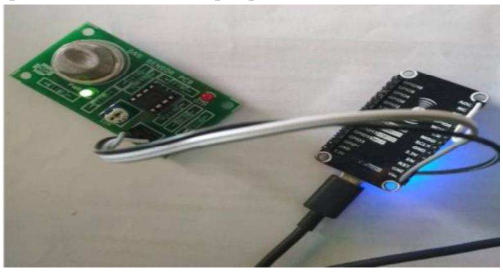


Fig 3: Data collection from sensor

4. Pre-Processing: Data Pre-processing is an important step during knowledge discovering. The collected data may contain missing, fault and outliers etc., Removal of these kinds of invalid data may produce misleading outcomes and makes knowledge discovery a challenge. Data is pre-processed by different ways such as cleaning, normalization, transformation, feature extraction and selection, etc. The major obstacle with clinical data is that redundant records and these records are eliminated to enhance the detection accuracy. Data transformation and data validation are two important pre-processing techniques.

5. Data Storage: The data stored in a cloud storage system with remote servers that accessible by internet and it managed, operated, and maintained by service provider. This proposed approach, the collected data are stored in ThingSpeak which is a cloud service provider. The flow of storage is showed in the fig 4.

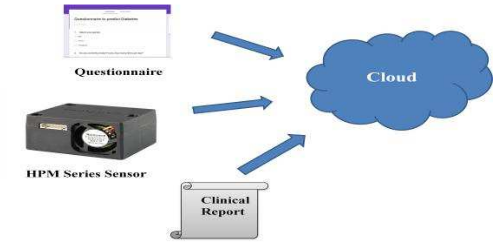


Fig 4: Collection of various data)

IV. PREDICTION OF DIABETES

The study made on various classification algorithms used in existing methods, three algorithms play major role in predicting Diabetes mellitus. They are J48, KNN, and Naïve Bayes. The PIMA Indian Dataset was applied to these 3 algorithms in which J48 algorithm predicts results with better accuracy [15]. So in this study J48 is used and the collected data is applied in WEKA to classify Diabetes Mellitus based on different attributes like age, sex, income, education, work type, blood pressure (diastolic and systolic), body mass index (BMI), dietary history, physical activity, pattern and Pm (Pm2.5& Pm10). The outcome of predicting Diabetes Mellitus is represented as a class variable 1 or 0, depending on whether the person has diabetes or not respectively.

The nature of the collected data has described in this section. The overall male and female from the total study population has been separated based on their age with a percentage of the population and it is listed below in the table 1.

Table 1: Distribution of population based on their age and sex

Age	No. Male Population (%)	No. Female Population (%)	Total Population (%)
< 30 years	42(59.15 %)	29(40.84%)	71 (5.81)
30- 35 years	31 (58.49%)	22(41.50)	53(4.34)
36- 40 years	172(64.66)	94 (35.33)	266(21.78)
41- 50 years	612 (48.84)	310 (51.15)	606 (49.63)
51- 60 years	118(68.20)	55 (31.79)	173 (14.16)
>60 years	21(40.38)	31(59.61)	52(4.25)

A. **Family and Income:** From the study of population, people are separated based on their family and income. They were grouped into four categories based on their income style such as below 50,000, 50,000 to 1,50,000, 1,50,000 to 2,00,000 and above 2,00,000. According to these categories, people were separated like diabetic and non-diabetic and tabulated as shown in table 2.

Table 2: population separated based their monthly income

Income	Total	Percentage of total (%)
Below 50,000	341	27.92
50,000 to 1,50,000	662	54.21
1,50,000 to 2,00,000	161	13.18
above 2,00,000	57	4.66

B. Education: In Tiruchirappalli district, people are living with various education levels, such as school, college, and illiterate. These survey details are given in the fig 5.

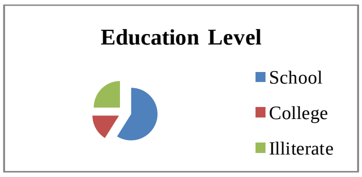


Fig 5: Education level based division

C. Work Type: According to the physical work of individuals, the work is categorized as easy, medium, and hard and based on their work type the details about diabetic patients were represented in the fig 6.

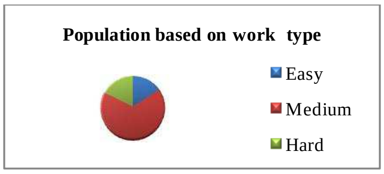


Fig 6: Population divided by work type.

D. Awareness of Diabetes Test: People who have diabetes are certainly aware of the disease and also will be aware of the precautions to be taken. The evaluation of awareness among people is depicted as a graph in fig 7.

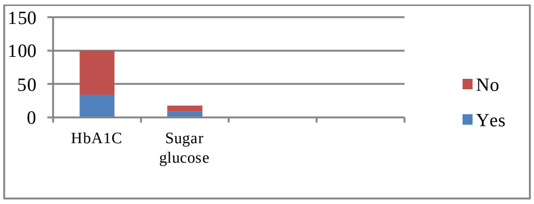


Fig 7: Awareness about Diabetes Mellitus

Furthermore, sugar count helps to find out the sugar level of an individual, suppose if a person has a sugar count below 140 then it is known as low sugar level, or if the sugar count is above 140 to 180 then the sugar level is normal, which is also called as pre-diabetic but if the sugar count exceed above 180 then the count is high. The surveyed result is shown in Table3.

Table 3: Sugar level based on the sugar test.

	low sugar	pre-diabetes	high sugar
below 140	37.2		
140 - 180		42.6	
above 180			20.2

E. Blood Pressure and Work Type: Blood pressure varies based on the people's work type. There are three categories of works such as easy, medium and hard. The pressure level is also divided into high, medium and normal. Figure 8 depicts the list of people who have blood pressure, which is separated based on easy, medium and hard type of work.

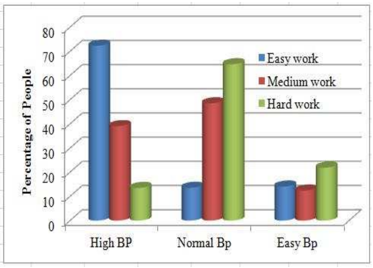


Fig 8: Work type vs Blood pressure level

F. Smoking and Liquor Drinking Habits: People, who are smoking, consuming alcohol, both smoking & consuming alcohol are 314, 193 and 178 respectively

Table4: List of data related with smoking and drinking.

7. Air Quality: Air quality is as an important factor in this study because it also one of the reason for diabetes mellitus. The air quality level is measured through the $PM_{2.5}$ and PM_{10} level in the air and fixed into the area to evaluate the particle level. From this the PM level is measured and separated among diabetes people that showed in table 5.

Table 5: Air quality and Diabetes

Air Quality	Diabetic	Non-Diabetic
High	68	36
Medium	15	47
Low	17	17

V. CONCLUSION

In Machine Learning data patterns are extracted by applying intelligent methods. These methods provided the great opportunities to assist physicians deal with this large amount of data. This study provided a view about the prevalence of diabetes mellitus using classification techniques. It helps the patients to prevent themselves from the disease. Decision tree model has outperformed than naïve Bayes and KNN techniques. The proposed work detected that the polluted air causes the diabetes and also increases the risk of diabetes. The proposed work can be further enhanced and expanded with stacking techniques to increase the accuracy of prediction..

ACKNOWLEDGMENT

This research work is financially supported by University Grants Commission, Government of India, under the Minor Research Project scheme. Ref. No.: F MRF-6517/16 (SER)/UGC).

REFERENCES

1. Arun K Pujari, "Data Mining Techniques", Universities Press (India) Private Limited2001
2. Krochmal, Magdalena, and Holger Husi, "Knowledge discovery and data mining" Integration of Omics Approaches and Systems Biology for Clinical Applications , 2018, pp. 233-247.
3. Qi Luo. "Advancing Knowledge Discovery and Data Mining", IEEE Workshop on Knowledge Discovery and Data Mining,2008.
4. S.D.Gheware, A.S.K. ejkar,S.M.Tondare,"Data mining: Task, Tools techniques and applications", International Journal of Advanced Research in Computer and Communication Engineering, Vol.3, Issue.10,2014, pp. 8095 -8098.
5. Krishnaiah, V. Narsimha, G. and Subhash Chandra, N. "A Study on Clinical Prediction using Data Mining Techniques", International Journal ofComputer Science Engineering and Information Technology Research (IJCSEITR), Vol.3, Issue.1, 2013, pp.239-248.
6. Himansu Das, Bighnaraj Naik and H. S. Behera,"Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", Springer Nature Singapore Pte Ltd,2018,pp:539-549.
7. Miss. N. Vijayalakshmi, Miss. T. Jenifer,"An Analysis of Risk Factors for Diabetes Using Data Mining Approach",International Journal of Computer Science and Mobile Computing, Vol.6, Issue.7, 2017, pp:166 – 172.
8. Kalaiselvi, C., and G. M. Nasira. "Prediction of heart diseases and cancer in diabetic patients using data mining techniques." Indian Journal of Science and Technology ,Vol.8, Issue. 14 ,2015.
9. Swaroopa Shastri, Surekha, Sarita, " Data Mining Techniques to Predict Diabetes Influenced Kidney Disease", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol.2, Issue. 4, 2017, pp. 364-368.
10. Chang, Huan-Cheng, Pin-Hsiang Chang, Sung-Chin Tseng, Chi-Chang Chang, and Yen-Chiao Lu. "A comparative analysis of data mining techniques for prediction of postprandial blood glucose: A cohort study." International Journal of Management, Economics and Social Sciences (IJMESS) , Vol.7 ,2018, pp. 132-141.
11. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, Issue.1, 2015, pp. 1-14.
12. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes healthcare in young and old patients", Journal of King Saud University - Computer and Information Sciences, 2013, Vol.25, pp. 127-136.
13. Emad A Mohammed, Behrouz H Far and Christopher Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends", BioData Mining 2014, Vol.7, pp.1-23, http://www.biodatamining.org/content/7/1/22
14. Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S, " Predictive Methodology for Diabetic Data Analysis in Big Data", Procedia Computer Science 50, 2015, pp. 203 - 208, Available online at www.sciencedirect.com.
15. Dr. L. Arockiam, A. Dalvin Vinoth Kumar, S. Sathyapriya, "Performance Analysis of classification Algorithms for Diabetic Prediction Using Pima- Indian dataset", Journal of Emerging Technologies and Innovative Research (JETIR), Vol. 5, No.12, 2018, pp.563-569.

AUTHORS PROFILE

First Author Dr. L. Arockiam is working as Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Thiruchirapalli, Tamil Nadu, India. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.

Second Author S.Sathyapriya is doing her Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. Her research area is IoT Data Analytics.

Third Author V. A. Jane is doing his Ph.D in Computer Science in St.Joseph's College (Autonomous), Thiruchirapalli, Tamilnadu, India. His research area is IoT Data Analytics.

Fourth Author Dr. A. DalvinVinoth Kumar is working as Assistant Professor in the Department of Computer Science, Kristu Jayanthi College Bengaluru, Karnataka, India. His research interests are: MANET, Routing and IoT

Source details

International Journal of Innovative Technology and Exploring Engineering

Scopus coverage years: from 2018 to 2019
(coverage discontinued in Scopus)
Publisher: Blue Eyes Intelligence Engineering and Sciences Publication
E-ISSN: 2278-3075
Subject area: Engineering: Civil and Structural Engineering | Computer Science: General Computer Science | Engineering: Electrical and Electronic Engineering | Engineering: Mechanics of Materials
Source type: Journal

CiteScore 2019
0.6

SJR 2019
0.102

SNIP 2020
0.387

View all documents > | Set document alert | Save to source list | Source Homepage

CiteScore | CiteScore rank & trend | Scopus content coverage

i Improved CiteScore methodology
CiteScore 2019 counts the citations received in 2016-2019 to articles, reviews, conference papers, book chapters and data papers published in 2016-2019, and divides this by the number of publications published in 2016-2019. Learn more >

CiteScore 2019

0.6 = 5,200 Citations 2016 - 2019 / 8,638 Documents 2016 - 2019

Calculated on 06 May, 2020

CiteScore rank 2019

Category	Rank	Percentile
Engineering — Civil and Structural Engineering	#249/310	19th
Computer Science — General Computer Science	#187/221	15th

View CiteScore methodology > CiteScore FAQ > Add CiteScore to your site

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8, Issue-11, September 2019

Nutrition Monitoring and Calorie Estimation using Internet of Things (IoT)

P. Kamakshi Priyaa, S. Sathyapriya, L. Arockiam

Abstract*: Diet observation is one of the principal aspect in precautionary health care that aims to cut back varied health risks. The various recent advancements in smartphone and wearable sensing element technologies have paved way to a proliferation of food observation applications that are based on automated image processing and intake detection, with an aim to beat drawbacks of the standard manual food journaling that's time overwhelming, inaccurate, underreporting, and low adherent. The currently developed food logging methods are very much time consuming and inconvenient that limits their effectiveness. The proposed work presents an Internet of Things (IoT) based mobile-controlled calorie estimation system to make technical advancements in healthcare industry. The proposed system operates on mobile environment, which allow the user to acquire the food image and quantify the calorie intake mechanically. The Mqtt protocol based MyMqtt broker is used to connect the application and the edge device and also to store the data in the Thingspeak cloud. A deep convolutional network is employed to classify the food accurately within the system. The volume estimation is done using sensors and the calorie approximation is done using formula.

***Keywords* : Calorie Estimation, Convolutional Neural Network, , Deep Learning, IoT.**

I. INTRODUCTION

A well-balanced healthy diet is a vital goal for several individuals. A technique to attain this is by manually recording the quantity of calories consumed. This pursuit method, however, is often terribly tedious because it needs the user to have a food journal and to try and do complex calculations to be able to estimate the quantity of calories consumed in each food item. In fact, it has been conjointly shown that individuals tend to underestimate the amount of calories most of the time [1]. Nutritional imbalance will occur either because of under-nourishment or over-nourishment. Under-nutrition is a condition where the sufficient nutrients are not consumed. Malnourishment will result in organic process disorders which may be characterised by symptoms like anxiety, mood swings, tender bones, injury gums, and skinny hairline etc. Similarly, over-nutrition is a condition where multiple varieties of imbalanced food are consumed. Over-nutrition will increase the chance of obesity that is primarily caused by the intake of food that is high in salt and fat. The Internet of Things (IoT) aids in connecting real world sensing information to predominantly cloud based elucidations.

Revised Manuscript Received on September 2, 2019.

S. Sathyapriya, Ph.D. Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

Dr. L. Arockiam, Associate Professor Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

It is an associate internetwork of sensors deployed within the physical world and helps in swapping data between these sensors and also the cloud [2]. In the current years, there are a number of measures supported by computer vision to estimate calories [3, 4]. For these strategies, the accuracy of estimation results are determined by two main factors namely: object detection algorithms and volume estimation methodology. Within the side of object detection, classification algorithm like Support Vector Machine (SVM) is used to recognise the generic food items. With the aspect of volume estimation, the standardisation of food and also the volume calculation are the key problems. The appliance of deep learning for food classification and recognition is extremely recommendable. Deep learning is associate rising approach from machine learning, and has been projected in recent years to switch the machine learning systems towards the discovery of multiple levels of illustration. The recent advancements in this area was surveyed in a detailed manner in [5]. The proposed system uses a robust deep learning based neural network model as a way of improvising the reliability of food image classification and calorie estimation systems. The model integrates the mobile calorie activity application to the deep neural network. The Convolutional Neural Network (CNN) is a pillar that handles the training and testing requests at the upper layers without affecting the central layers.

II. RELATED WORKS

An innumerable amount of techniques have been applied to measure the calorie within the food. One amongst the primary clinical works in the domain is the 24 Hour Dietary Recall (24HR) [6], [7]. This method records the daily calorie intake limited to a duration of twenty four hours. The user is anticipated to recall the food items that have been consumed in the previous twenty four hours prior to the interview. In this technique, the estimation of food portion size is done using standardized cups and spoons. Based on the measure of food, the nutrient consumed is estimated using the food composition tables. Food Frequency Questionnaire (FFQ) is a technique [8] that focuses on elaborating the dietary patterns or food habits, however not the calorie intake. The major limitation of the 24HR and FFQ are namely the delay time in outlining the food consumed, the under-reporting of the food portion size, human memory dependency, requirement of qualified interviewers who can predict the quantity of calories and nutrients and the requirement of performing complicated calculations to estimate the frequencies.

The process of automatically predicting the quantity of calories in food using their images has been receiving some attention within the domain of computer vision. In [9] a food image recognition system was presented which was further enhanced to restaurant specific food logging system in [10]. A mobile phone based food classification system which assess the nutrient intake has been presented in [11]. The wearable sensor based automated methods could be a potential solution as they require minimal user intervention. These sensors adopt various detection techniques that are primarily based on the user activities [12]. An automated nutrition prediction monitoring system was proposed using machine learning techniques in [13] which was further improvised as a deep learning based model in the subsequent research [14]. There are numerous research works that have been performed to adopt the sensors to monitor food intake, but most of them are suited only for laboratory conditions along with limited participants. The average accuracy of the systems engineered using these sensors is about 90% [15]. So, it is a challenge to find an efficient solution to monitor the nutritional consumption with greater accuracy in the real-world context.

III. PROPOSED APPROACH

The proposed work aims to build an IoT Based Automated Nutrition Monitoring and Calorie Estimation System in mobile environment. The users of the system have the mobile interface through which the demographic information and physical measurement data are collected. The application is used to activate the microcontroller module through the MyMqtt broker. The microcontroller triggers the USB camera to capture a photo of the food item placed on the load sensor and the weight of the timestamped food item is also estimated. The collected data are stored in the Thingspeak cloud server. The subsequent data analytics is done using the collected data. Initially, a deep learning based Convolutional Neural Network model is built based on the classification of food items. The developed model is then used to predict the type of food. Based on the volume estimated by the load sensor and the prediction of the food item, the USDA (US Department of Agriculture) standardized data repository is utilized for collecting the nutritional facts which in turn determines the amount of calorie that was consumed by an individual. The architecture of the proposed system has been depicted pictorially in Fig. 1 and there is a subsequent elaborated view of all the activities in the system.

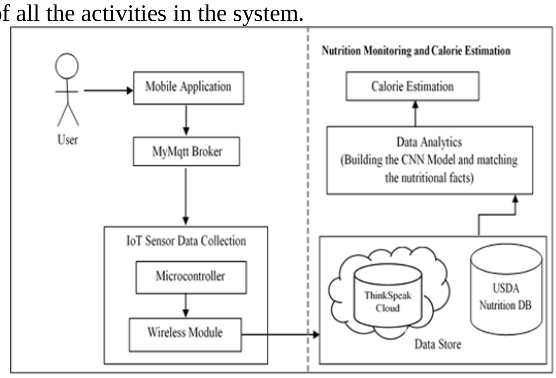


Fig. 1. Proposed Architecture

A. IoT Sensor Data Collection

This is the interface between the user and the system which collects the data using the data collection points. The user display the food item to the sensing element system and authenticate it via a mobile app. The mobile application publishes the request to activate the sensor system by the MyMqtt Broker. The microcontroller receives the request as it has subscribed to the mutual topic. The microcontroller activates the USB camera which captures the photo of the food item. The activated load sensor estimates the weight of the food placed on the weighing sensor along with the timestamp. The working procedure is represented in Fig. 2.

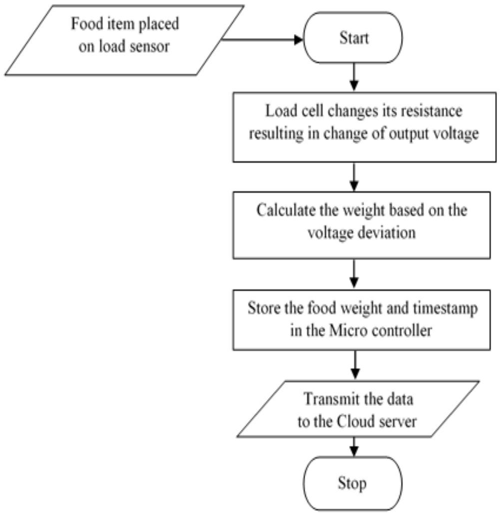


Fig. 2. IoT Data Flow

The experimental setup is shown in Fig. 3.

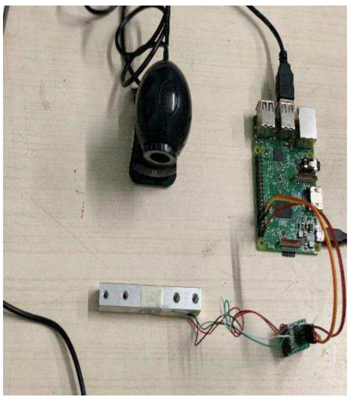


Fig. 3 Experimental Setup

B. Datastore

The data store contains both the raw data (sourced from the sensors and the mobile application) and the nutritional facts data (sourced from the USDA API).

C. Cloud Servers

The data received from the sensors and camera are temporarily stored in local SD card in the microcontroller and are later transmitted to the Thingspeak Cloud server.

The Thingspeak cloud server is employed for remote storage purpose. The data stored in the Thingspeak cloud is used for the analytics process.

D. Data Analytics

The Data Analytics module is responsible for statistical analysis and deep learning events in the system. It includes algorithms to detect objects, build a CNN model, estimate volume and associated nutritionary composition values and integrate the information. Fig. 4 displays a portion of the dataset used for model evaluation. Fig. 4 displays the portion of the image dataset.

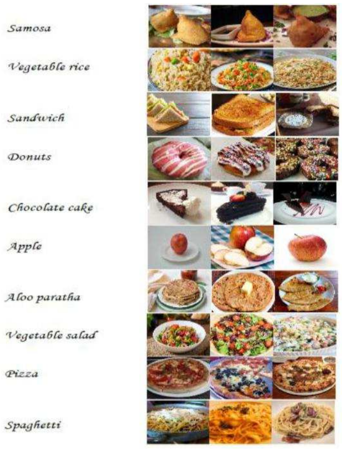


Fig. 4. Sample Dataset

- **Building the model**

The process of building the deep neural network model is done when the data acquisition is done. The initial step for building the model is to produce a pre-trained model file using the CNN network. The model is built by collecting the images of a particular class (approximately 1000 images per class) and then using the object name-set they are labelled. The collected images are used to train the model. The model file thus generated is used to perform image recognition against the image captured by the user. The system predicts a list of probable class labels for the image and the class with the highest probability is assigned.

The various steps involved in building the CNN model is summarized as follows.

Step 1: The initial setup which comprises of importing the necessary packages namely Sequential, Convolution2D, MaxPooling2D, Flatten and Dense.

Step 2: The neural network is initialized using an object.

Step 3: The convolutional layers are added specifying the filter size, shape, dimensions and the activation function.

Step 4: Pooling is performed for feature map size reduction without losing the important image characteristics.

Step 5: Flattening is done by collecting all the pooled feature maps and putting them into a single vector.

Step 6: The full connection is created using the number of hidden layer nodes, relu function for hidden layer, softmax function for multiple outcomes and the output probabilities are randomly generated.

Step 7: The CNN model is compiled and image augmentation is done to train and test the model.

Step 8: The model is tested with unseen (new) image and the output is evaluated in terms of classification accuracy and Mean Square Error (MSE).

The constructed model's accuracy for 25 epochs is displayed in Fig. 5 and Fig. 6 shows the model's loss.

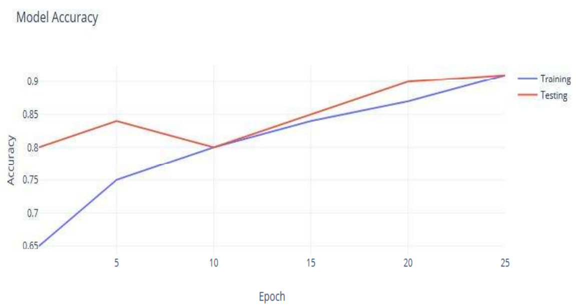


Fig. 5. The accuracy of the model with 25 epochs

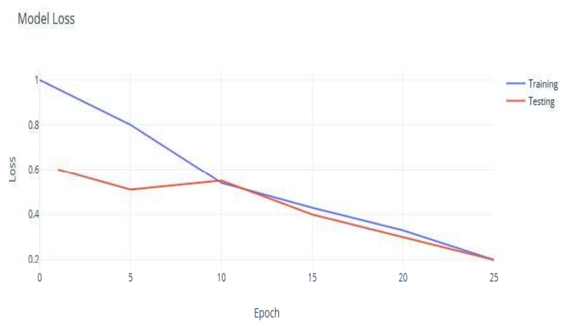


Fig. 6. The loss of the model with 25 epochs

E. Calorie Estimation

The prototype consists of the Weighing Load Cell Sensor (5kg) connected with a Raspberry Pi 3 microcontroller which can be efficiently programmed using Python. The data logging is achieved through the USB port. An approximate estimation of the weight is provided by the sensor (in gm). Once the weight is estimated and the food label is predicted, it can be utilized to calculate the calorific value using the formula. The SR Legacy database available at the US Department of Agriculture website is used to estimate the calorific value.

$$\text{Calorie Value} = \frac{\text{Calorific table value} * \text{Weight of the predicted food}}{\text{Mass in the Nutrition Table}}$$

A sample of the estimated calorie value using the prediction model is provided in Table-I.

Table- I: Calorie Estimation

Food item	Standard Calorie (100 gm)	Mass obtained (in gm)	Estimated Calorie
Samosa	262	107	280.34
Vegetable rice	129	256	330.24
Sandwich	243	308	748.44
Donuts	396	43	170.28
Chocolate cake	458	86	393.88
Apple	254	62	157.48
Aloo paratha	240	273	655.20
Vegetable salad	170	208	353.60
Pizza	269	363	976.47
Spaghetti	66.64	564	375.85

IV. CONCLUSION

In the proposed work, a food data logging system which estimates the calories using IoT has been developed. This system analyzes the average calorie intake of a person in a day and helps them by providing some insights about their diet. The study utilizes a cost effective sensor system and seamless data logging method, which makes the system more desirable. The deep learning model achieved an accuracy of 90.69% with a loss of 0.2324. The future work is to cover more food types from a variety of cuisines around the world. In addition, an obesity prediction system considering the daily calorific intake along with other relevant attributes is to be developed to analyze the correlation among obesity and the attributes.

REFERENCES

1. Elbel B, "Consumer estimation of recommended and actual calories at fast food restaurants", Obesity, Vol. 19, Issue 10, 2011, pp. 1971 – 1978.
2. Kougianos E, Mohanty S P, Coelho G, Albalawi U and Sundaravadivel P, "Design of a High-Performance System for Secure Image Communication in the Internet of Things," IEEE Access, Vol. 4, 2016, pp. 1222 – 1242.
3. Jia W, Chen H C, Yue Y, Li Z, Fernstrom J, Bai Y, Li C and Sun M, "Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera", Public Health Nutrition, Vol. 17, Issue 8, 2014, pp. 1671 – 1681.
4. Pouladzadeh P, Kuhad P, Peddi S V B, Yassine A and Shirmohammadi S, "Mobile cloud based food calorie measurement", In Proceedings of the International Instrumentation and Measurement Technology, 2016, pp. 1 – 6.
5. Kamakshi Priyaa P, Arockiam L, "A Survey on Nutrition Monitoring and Dietary Management System", IJRAR- International Journal of Research and Analytical Reviews (IJRAR), Vol. 6, Issue 2, 2019, 329-334, DOI: http://doi.one/10.1729/Journal.21493.
6. Livingstone M B E, Robson P J and Wallace J M W, "Issues in dietary intake assessment of children and adolescents", British Journal of Nutrition, Vol. 92, Issue S2, 2004, pp. S213 – S222.
7. Bandini L, Must A, Cyr H, Anderson S, Spadano J and Dietz W, "Longitudinal changes in the accuracy of reported energy intake in girls 10-15 y of age", The American Journal of Clinical Nutrition, Vol. 78, Issue 3, 2003, pp. 480 – 484.
8. Luo W, Morrison H, de Groh M, Waters C, DesMeules M, Jones-McLean E, Ugnat A M, Desjardins S, Lim M and Mao Y, "The burden of adult obesity in Canada", Chronic Diseases in Canada, Vol. 27, Issue 4, 2007, pp. 135 – 144.
9. Chen M Y, Yang Y H, Ho C J, Wang S H, Liu S M, Chang E, Yeh C H and Ouhyoung M "Automatic Chinese Food Identification and Quantity Estimation," In SIGGRAPH Asia 2012 Technical Briefs, ACM, 2012.
10. Oscar Beijbom , Neel Joshi , Dan Morris , Scott Saponas and Siddharth Khullar, "Menu-Match: Restaurant-Specific Food Logging from Images", Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 844 – 851, [doi>10.1109/WACV.2015.117].
11. Kiran Abhore and Dawande N A, "Measuring Calories and Nutrition from Food Image", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 6, 2016.
12. Niloofar Hezarjaribi, Cody A, Reynolds, Drew T. Miller, Naomi Chaytor and Hassan Ghasemzadeh, "S2NI: A mobile platform for nutrition monitoring from spoken data", 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 1991 – 1994.
13. Prabha Sundaravadivel, Kavya Kesavan, Lokeshwar Kesavan, Saraju P. Mohanty, Elias Kougianos and Madhavi Ganapathiraju, "Smart-log: An automated, predictive nutrition monitoring system for infants through the IoT", IEEE International Conference on Consumer Electronics (ICCE), 2018.
14. Prabha Sundaravadivel, Kavya Kesavan, Lokeshwar Kesavan, Saraju P. Mohanty and Elias Kougianos, "Smart-Log: A Deep-Learning based Automated Nutrition Monitoring System in the IoT", IEEE Transactions on Consumer Electronics, Vol. 64, Issue 3, 2018, pp. 390 – 398.
15. Vu T, Lin F, Alshurafa N and Xu W, "Wearable food intake monitoring technologies: A comprehensive review", Computers, Vol. 6, Issue 1, 2017.

AUTHORS PROFILE

P. Kamakshi Priyaa is pursuing her M.Phil. in Computer Science at St .Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. Her research area is IoT Data Analytics.

S. Sathyapriya is doing her Ph.D. in Computer Science at St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. Her research area is IoT Data Analytics.

Dr. L. Arockiam is working as an Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.

Scopus Preview   Author search   Sources   Create account   Sign in

Source details

International Journal of Scientific and Technology Research

Scopus coverage years: from 2018 to 2020

(coverage discontinued in Scopus)

Publisher: International Journal of Scientific and Technology Research

E-ISSN: 2277-8616

Subject area: Social Sciences: Development   Business, Management and Accounting: Management of Technology and Innovation   Engineering: General Engineering

Source type: Journal

View all documents >   Set document alert   Save to source list   Source Homepage

CiteScore 2019
0.2

SJR 2019
0.123

SNIP 2020
0.485

CiteScore   CiteScore rank & trend   Scopus content coverage

i Improved CiteScore methodology

CiteScore 2019 counts the citations received in 2016-2019 to articles, reviews, conference papers, book chapters and data papers published in 2016-2019, and divides this by the number of publications published in 2016-2019. Learn more >

CiteScore 2019

$$0.2 = \frac{\text{679 Citations 2016 - 2019}}{\text{3,542 Documents 2016 - 2019}}$$

Calculated on 06 May, 2020

CiteScore rank 2019

Category	Rank	Percentile
Social Sciences └ Development	#215/239	10th
Business, Management and Accounting └ Management of Technology and Innovation	#209/231	9th

View CiteScore methodology >   CiteScore FAQ >   Add CiteScore to your site

An Enhanced Decision Tree Ensemble Technique For Obesity Prediction

P. Kamakshi Priyaa, S. Sathyapriya, Dr. L. Arockiam

Abstract: Currently, obesity which is a non-communicable disease (NCD) is a serious health issue that leads to many life-threatening diseases like diabetes (Type 2), cancer and heart ailments. Obesity can be determined using various factors, namely age, weight, height and Body Mass Index (BMI). The techniques that are used for obesity prediction are extremely reliant on the BMI; however, the BMI cannot be applied universally. The proposed R ensemble based prediction model incorporating 13 variables has provided an average accuracy of 97.29%. The ensemble model leverages the Enhanced Decision Tree, Naïve Bayes and the Support Vector Machine (Radial SVM). The work has considered unique health parameters such as calories consumed, pregnancy status and bodybuilder which have proved to drastically improve the accuracy of the prediction model.

Index Terms: Ensemble Learning, Machine Learning, Non-Communicable disease, Obesity, Prediction.

1 INTRODUCTION

Obesity is becoming one of the most serious global health problem. Obesity is a term coined for the anatomical condition characterized by an excessive growth of body fat, individually the build-up of adipose tissue beneath the skin [1]. The adults and elderly people affected by obesity are at a higher risk to attract cardiovascular ailments [2] and the people who tend to be overweight tend to be affected by serious health issues in their latter part of life. Obesity is a prominent factor which gives rise to diseases like type 2 diabetes, high blood cholesterol, high blood pressure, heart ailments hypertension, dyslipidaemia and metabolic syndrome [3], [4]. The amount of people diagnosed with clinical obesity has increased greatly in recent years [5]. The psychological, physical and economic consequences of obesity has been discussed well in [6]. The intake of food items rich in saturated fat is directly associated with obesity [7], leading to recommendations to reduce body weight. The inference from studies shows that the process of weight loss has proven to make moderate change in reducing the risk factor among groups and provides a considerable amount of health benefits. Many people around the globe are ignorant and they are not conscious of maintaining a healthy body which reduces the risk of developing many diseases. Machine learning represents a broad array of different techniques, which can be broadly grouped based on the learning methods. The supervised methods include classifiers, the unsupervised methods include clustering and the semi-supervised methods include label propagation. The supervised machine learning technique has been incorporated in this research work.

- *P. Kamakshi Priyaa is pursuing her M.Phil. in Computer Science at St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: kamakshi1796@gmail.com. Her research area is IoT Data Analytics.*
- *S. Sathyapriya is doing her Ph.D. in Computer Science at St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: sathyapriya2822@gmail.com. Her research area is IoT Data Analytics.*
- *Dr. L. Arockiam is working as an Associate Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India, E-mail: larockiam@yahoo.co.in. His research interests are: Software Measurement, Cognitive Aspects in Programming, Data Mining, Mobile Networks, IoT and Cloud Computing.*

Machine learning approaches are being widely used to predict the risk of obesity by identifying causes and developing proactive strategies to prevent it. In the recent times many extensive factors are being considered as possible influencers of obesity.

2 RELATED WORKS

There are several research works which have been explored using data mining techniques in health and disease prediction [8], [9]. Machine learning and structural equation modelling techniques are being used to examine large amounts of data to identify patterns and relationships that would otherwise go undetected [10], [11]. Some studies have applied these methods to build predictive models to understand the obesity problem. The tendency of obesity was examined within selected 7 English medium school students and most of them consume the injurious fast food from the road side shop and get obese [12]. The author designed a cross-sectional study by using a statistical software SPSS. Meng et al., [13] proposed a system that compared the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes using common risk factors. The logistic regression model gave the highest accuracy and the ANN gave the lowest accuracy. The general attributes namely the Body Mass Index (BMI), age, physician supply, ethnicity and education are utilized for mining the clinical data to predict the heart disease. The BMI does not match the risk factor condition. Some of the risk factors like smoking and regular intake of sugar-sweeten beverage increased the obesity risk to a greater extent than other factors in the daily life [14], in which smoking was not a statistically significant risk factor in [15]. Verma et al., [16] introduced a Multi-layer Perceptron algorithm to achieve higher level prediction accuracy of 88.4%. The proposed hybrid model improves the class level accuracy for the coronary artery disease. Nur' Aina Daud et al., [17] proposed a system to predict obesity based on grocery data using J48 algorithm with an accuracy of 89.4118%.The dataset was based on three different types of data that were manually collected namely grocery data (food weight, quantity, description), demographic data (age, gender) and anthropometric data containing body height, weight and physical activity level. The work was limited to a compatibility of only one algorithm and also the calorie value was assigned based on approximation formula. Given the shortcomings of previous research, the present study aims to build an ensemble based machine learning model to predict obesity.

3 PROPOSED APPROACH

The proposed work aims to build an Ensemble based Obesity Prediction Model using machine learning approaches in mobile environment. The initial step is the data collection which is conjointly done using a mobile application and also through questionnaires distributed through Google Forms. The data collected from these dual sources are stored in the data storage repository. Data Preprocessing is done for feature selection using correlation analysis and to discard the missing values using the median value and the preprocessed data are stored in the data storage. The ensemble based model using Stacking method is built by analyzing the performance of 3 machine learning algorithms namely Enhanced Decision Tree, Naïve Bayes and Support Vector Machine (Radial SVM).Finally, obesity prediction is done by the top layer model which is the most efficient one and the performance of the model is evaluated. The architecture of the proposed system is shown in Fig. 1. A brief description of the various activities in the system are summarized in the following section.

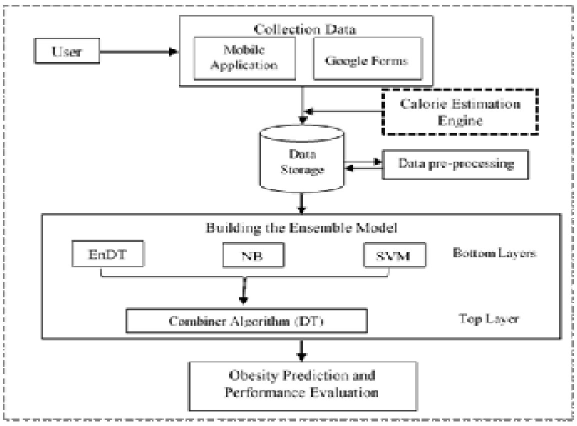


Fig. 1 *Proposed Architecture*

3.1 Data Collection

The data is collected from two sources namely through a mobile application and questionnaires distributed through Google Forms. Given below is a detailed overview of the process of data collection.

3.1.1 Mobile Application

An android based mobile application was developed using Android Studio which acts as the interface between the users and the system. The mobile application included the following set of attributes that can be split into 3 categories.

- Category A: contains demographic information including age and gender.
- Category B: contains the physical data like height, weight, waistline, physical activity level, and pregnancy status.
- Category C: contains the average calorie intake and frequency of processed food intake.

3.1.2 Questionnaire

The questionnaire was developed using Google Forms for the three categories of data as done through the mobile application. Unlike the mobile application which estimated the calories based on sensors, the questionnaire collected the calorie intake from the user. Based on the collected data, the following measures are calculated as

$$\text{BMI} = \frac{Weight\ in\ kg}{Height\ in\ m^2}$$

$$\text{WHfR} = \frac{Waist\ circumference\ (in\ cm)}{Height\ (in\ cm)}$$

The obesity index is calculated based on the values from Table 1. The average calorie intake was computed using a Deep learning based food image recognition system which is done in the previous work [18].

TABLE 1

BMI Based Obesity Chart

BMI (kg/m^2)	*Weight Classification*	*Obesity class*
<18.5	*Underweight*	-
18.5-24.9	*Normal*	-
25.0-29.9	*Overweight*	-
30.0-34.9	*Obese*	*Obesity class 1*
35.0-39.9	*Highly Obese*	*Obesity class 2*
>= 40	*Extremely Obese*	*Obesity class 3*

3.2 Data Storage

The data collected from both the sources were combined together and stored in .csv (Comma Separated Values) format. The collected data accounted to 324 records. The datatypes of the attributes were modified in order to suit the needs of the machine learning algorithm. The list of attributes and their description is provided in Table 2. These data were utilized to build the model.

TABLE 2

ATTRIBUTE AND DESCRIPTION

Attribute Name	*Description*
Gender	*Contains the gender of the person. Males are denoted as 0 and females as 1*
Height	*Numeric value that contains the height of an individual expressed in cm*
Weight	*The field contains the weight of the person expressed in kg*
Age	*The age of an individual*
Calorie_Intake	*The average amount of calories the user intakes per day expressed in kcal*
Waistline	*The circumference of the waistline was obtained in inches and was converted into cm.*
Pregnency	*The field determines whether a woman is pregnant which has values of 0 denoting no and 1 denoting yes.*
Physical_Activity_Level	*The attribute denotes the routine physical activity level of an individual which falls in either one of the three categories namely Extremely Active denoted as 0, Moderately Active denoted as 1 and Sedentary as 2.*
Processed_Food_Intake	*The field displays the frequency of processed food intake by the individual which was categorized as Rarely denoted as 0, Occasionally denoted as 1 and Frequently denoted as 2.*
BMI	*The BMI is calculated using the formula (weight (kg)/height (m^2) and the computed value is rounded to one decimal place*
Index	*The attribute denotes the BMI Index which is calculated from the standard BMI table which contains levels from 1 to 6*
WHfR	*The waist to height ratio is calculated by weight (cm)/height (cm)*
Obesity_Predict	*The target attribute which predicts whether or not an individual has obesity which is denoted as 0 for no and 1 for yes.*

3.3 Data Storage

Data Pre-processing is an important step in the data mining process. The data that is collected may contain out-of-range values, missing values. The reliability of the data is increased when these kinds of invalid data are handled efficiently else it may lead to undesirable outcomes and makes the knowledge discovery process a challenge. The proposed system's dataset contained some missing values namely the waist size and processed food intake which were pre-processed using the median technique which generated the probable value. The waist size was calculated using the standard charts depending upon the height and weight. Feature selection was also done using the Correlational Attribute Evaluator and the resulting pre-processed data is sent for further processing.

3.4 Building the Ensemble Model

Ensembling is a technique of collaboratively working on more than two similar or dissimilar types of algorithms which are called the base learners. The process is done to build a more robust system which incorporates the predictions from all the base learners into account while making the final decision. This makes the final decision more robust, accurate and unlikely to be biased. The proposed R ensemble model is built using the stacking technique with two levels. The base level is formed using three supervised machine learning based algorithms namely Enhanced decision tree, Naïve Bayes and Support Vector Machine (Radial SVM). The algorithms were chosen such that they satisfy the requirement criteria of Stacking technique which includes that the individual model should have average accuracy criteria and the predictions of the individual models have low correlation with the predictions of other models. The model was trained with a p value of 0.50. A detailed description of the algorithms that were used to build the Ensemble Model are described below.

3.4.1 Enhanced Decision Tree

The decision tree model was chosen to be enhanced as the obesity dataset was highly compactable with the tree structure based algorithm and it was comprehensive and easy to interpret. The construction of a decision tree model involves a collection of decision nodes, connected by branches, extending downward from the root nodes until terminating in leaf nodes. The proposed algorithm provides an optimal enhancement of selecting the root node based on the correlation value of the attributes to the decision tree algorithm. The pseudocode of the modified algorithm is given below.

Pseudocode

Algorithm: DT_Learn (TT, TC, Can_Attr)

Input:

TT: set of training tuples

TC: the target class

Can_Attr: set of candidate attributes

Output:

A decision tree

Steps:

1) {
2) Create a root node RN //unlabeled node
3) if all the rows in TT are of the same target class T then
4) return RN as the leaf of the single node tree labeled with the class T;
5) if Can_Attr = empty (i.e. if no candidate attribute is present) then
6) return RN as the leaf of the single node tree labeled with the widely held value of the target in TT
7) Otherwise
8) {
9) Select the attribute Best_attri from Can_Attr that best classifies TT based on best positively correlated attribute
10) Set Best_Attri as the root node attribute
11) for each of the permitted value of Best_Attri, pv_i, //the tuples are partitioned and the subtrees are grown for each of the partition
12) {
13) Add a new branch below the root node that corresponds to Best_Attri=pv_i
14) let TT_{pvi} be the subset of TT that satisfies the outcome Best_Attri=pv_i
15) if TT_{pvi} = empty then

16) Add a leaf node which has the majority class in TT
    to the root node RN;
17) else
18) add the node that is returned by DT_Learn (TT_{pvi},
    TC, Can_Attr – {RN});
19) }
20) }
21) return (RN);
22) }

3.4.2 Naïve Bayes

The Naïve Bayes algorithm is simple yet an efficient probabilistic algorithm in classification technique which gets the probability value based on the frequency calculation and combinational values from the related collection. The algorithm assumes that the attributes are strongly independent. The inverse conditional probability is given by

$$P(Y|X_1\ldots\ldots X_n) = \frac{P(Y).P(X_1\ldots\ldots X_n|Y)}{P(X_1\ldots\ldots X_n)} \quad (1)$$

In equation (1), the variable Y is a class and the variable $X_1\ldots\ldots X_n$ denotes the classification characteristics. $P(Y|X_1\ldots\ldots X_n)$ is called the posterior probability and P(Y) is known as the prior probability. The conditional probability is computed as

$$P(X|Y = y) = \prod_{j=1}^{k} P(Xj|Y = y)$$

Pseudocode

Input:

Training dataset Tr,
$X = (X_1, X_2,\ldots\ldots X_n)$ //predictor variable value in the test dataset

Output:

A testing class dataset

Steps:

1) Read the training dataset T;
2) Calculate the mean value and the standard deviation value of the predictor variables in each class;
3) do
4) Calculate the probability of X_i using the gauss density equation in each class;
5) until
6) the probability of all the predictor variables (X_1, X_2,……X_n)
   has been calculated.
7) Calculate the likelihood of each class;
8) return the greatest likelihood;

3.4.3 Support Vector Machine

The support vector offer another approach to classify a multi-dimensional dataset in which the samples on the margin are called the support vectors. A SVM is a linear or non-linear classifier, which is mathematical function that distinguishes two different types of attributes.

Pseudocode

Input:

X and Y loaded with the training data, $\alpha \leftarrow 0$ or $\alpha \leftarrow$ partially trained SVM model

Output:

SVM classifier model

Steps:

1) C<= any value
2) Repeat
3) for all {Xi,Yi}, {Xj,Yj} do
4) Optimize α_i and α_j
5) end for
6) until no changes in α or other resource constraint criteria met
7) Ensure to retain only the support vectors ($\alpha_i > 0$)

3.5 Obesity Prediction

The obesity prediction in the proposed model is not completely dependent on the BMI alone. The WHfR is used so as a bodybuilder who has a higher BMI will have a low WHfR value. A pregnant lady may gain 11-15 kgs during pregnancy and so this factor was also considered. The average calorie intake per day based on the gender was included. The physical activity level and the frequency of processed food intake were also considered. In order to predict the obesity of an individual, all the attributes were considered which has proved to increase the accuracy of the proposed prediction model.

4 RESULTS AND DISCUSSION

This section presents the results that have been deduced from the proposed system. The prediction results of the base layer model containing Decision Trees with an accuracy of 96.87%, Naïve Bayes with an accuracy of 93.84% and svmRadial with an accuracy of 95.75% are shown in Fig. 2. The correlation between the various machine learning algorithms was analyzed and the maximum correlation of 0.41 was observed between decision tree and svmRadial which is less than the maximum threshold of 0.75 and hence the proposed set of algorithms are valid to be used for stacking. The plot matrix of the correlation between the models is shown in Fig. 3.

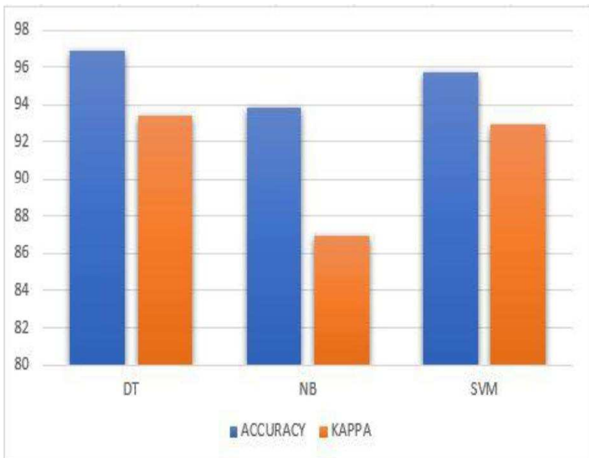


Fig. 2 *Comparison of Base Layer Accuracy*

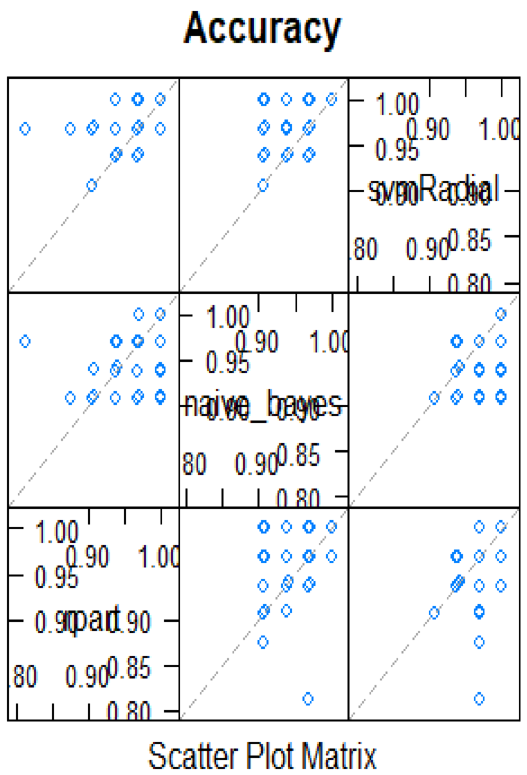


***Fig. 3**. Scatter Plot for model correlation*

The ensemble model with the decision tree as the top layer provided an increased accuracy of 97.29% which is an enhancement over the individual accuracy of 96.87%. A comparative analysis of the induvidual machine learning algorithm and the proposed ensemble model is shown in Fig. 4. The results prove that the proposed Ensemble model has improved the prediction accuracy.

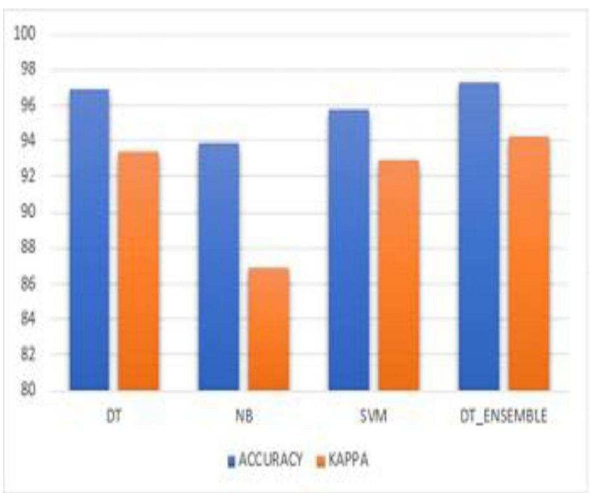


***Fig 4.** Comparative Analysis of Model Accuracy*

4.2 Performance Evaluation

The performance of the proposed system is compared with the related research area of obesity prediction and the proposed model has proven to have achieved a higher level of accuracy. Unlike the existing models which have only concentrated on the directly influential attributes of obesity, the proposed work has included extensive factors that have proven to have a positive correlation over the obesity level. The ensemble approach has proven to have a better prediction accuracy over the traditional algorithmic methods. The comparision of the previous research works is shown in Table 3.

TABLE 3

COMPARISON OF RELATED RESEARCH WORKS

Research Works	*Approach*	*Accuracy*
Zeyu et al. [19]	*Comparision of 4 Machine Learning Algorithms*	*88.82%*
Nur'Aina Daud et al. [17]	*Machine learning based J48 algorithm*	*89.41%*
Kapil et al. [20]	*Ensemble learning with three algorithms*	*89.68%*
Proposed work	*Ensemble learning with enhanced decision tree*	*97.29%*

The Table 3 shows that the proposed system has an improved accuracy over the previous related works. The comparative results show that ensemble models can be used to achieve a relatively higher accuracy than the individual algorithms.

5 CONCLUSION

The prediction of health conditions and diseases using machine learning techniques may be a challenging task but it increases the analytical accuracy and specificity. Data Analysis using machine learning techniques reduces the cost and time constraints involved. The proposed system has included the extensive factors of obesity which have proved to improve the prediction accuracy. The study has inferred that maintaining a healthy WHfR and an active physical activity level reduces the risk of obesity. The proposed ensemble model leveraging Enhanced Decision Tree, Naïve Bayes and Radial SVM provided an accuracy of 97.29%. The future work aims at including diverse factors for prediction and also increasing the levels of the ensemble model to further improve the accuracy of the model and also utilizing a larger dataset to analyze the model's efficiency.

6 REFERENCES

[1] N.L. Noor, N. Noordin, F.M. Saman and N.I. Teng, "Predicting Obesity from Grocery Data: A Conceptual Process Framework", In Proceedings of the 6th International Conference on Information and Communication Technology for the Muslim World (ICT4M), IEEE, pp. 286-291, 2016.

[2] M. Bouchonville, R. Armamento-Villareal, K. Shah, N. Napoli, D.R. Sinacore, C. Qualls and D.T. Villareal, "Weight loss, exercise or both and cardiometabolic risk factors in obese older adults: results of a randomized controlled trial", International Journal of Obesity, Vol. 38, No. 3, pp. 423, 2014.

[3] A. Dewan and M. Sharma, "Prediction of Heart Disease using a Hybrid Technique in Data Mining Classification", In Proceedings of the 2nd International Conference on

Computing for Sustainable Global Development (INDIACom), IEEE, pp. 704-706, 2015.
[4] J.D. Brown, J. Buscemi, V. Milsom, R. Malcolm and P.M. O'Neil, “Effects on Cardiovascular Risk Factors of Weight Losses Limited To 5–10%”, Translational Behavioral Medicine, Vol. 6, No. 3, pp. 339-346, 2015.
[5] S. Mitra, Y. Qiu, H. Moss, K. Li and S.L. Pallickara, “Effective Integration of Geotagged, Ancilliary Longitudinal Survey Datasets to Improve Adulthood Obesity Predictive Models”, In Proceedings of the International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, pp. 1738-1746, 2018.
[6] J.A. Gazmararian, D. Frisvold, K. Zhang and J.P. Koplan, “Obesity is associated with an increase in pharmaceutical expenses among university employees”, Journal of Obesity, 2015.
[7] R. Crescenzo, F. Bianco, A. Mazzoli, A. Giacco, R. Cancelliere, G. di Fabio, A. Zarrelli, G. Liverini and S. Iossa, “Fat quality influences the obesogenic effect of high fat diets”, Nutrients, Vol. 7, No. 11, pp. 9475–9491, 2015, DOI: https://doi.org/10.3390/nu7115480.
[8] P. Repalli, “Prediction on diabetes using data mining approach”, Oklahoma State University, 2011.
[9] N.A. Sundar, P.P. Latha and M.R. Chandra, “Performance analysis of classification data mining techniques over heart disease database”, International Journal of Engineering Science & Advanced Technology, Vol. 2, No. 3, pp. 470-478, 2012.
[10] J. Yang, H. Gu, X. Jiang, Q. Huang, X. Hu and X. Shen, “Walking in the PPI network to identify the risky SNP of osteoporosis with decision tree algorithm”, In Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1283-1287, 2016.
[11] T. Sivaranjani, “Comparative study on Obesity based on ID3 and KNN”, International Journal of Advance Research in Computer Science and Management Studies, Vol. 2, No. 9, pp. 389-396, 2014.
[12] M.N. Rahman, S.A. Reza, M.A. Islam, A. Rahman and A.K. Nath, “Prevalence of obesity and overweight among English medium school children of Dhaka City in Bangladesh”, Journal of Environmental Science and Natural Resources, Vol. 7, No. 1, pp. 63-67, 2014.
[13] X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang and Q. Liu, “Comparison of three data mining models for predicting diabetes or prediabetes by risk factors”, The Kaohsiung Journal of Medical Sciences, Vol. 29, No. 2, pp. 93-99, 2013.
[14] A. Pochini, Y. Wu and G. Hu, “Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents”, In Proceedings of the 3rd International Conference on Advanced Applied Informatics, IEEE, pp. 883-888, 2014.
[15] V. Silverwood, M. Blagojevic-Bucknall, C. Jinks, J.L. Jordan, J. Protheroe and K.P. Jordan, “Current evidence on risk factors for knee osteoarthritis in older adults: a systematic review and meta-analysis”, Osteoarthritis and Cartilage, Vol. 23, No. 4, pp. 507-515, 2015.
[16] L. Verma, S. Srivastava and P.C. Negi, “A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data”, Journal of medical systems, Vol. 40, No. 7, pp. 178, 2016.
[17] N.L. Nur, S.A. Aljunid, N. Noordin and N.I. Teng, “Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity”, In Proceedings of the IEEE Conference on Big Data and Analytics (ICBDA), IEEE, pp. 1-6, 2018.
[18] P. Kamakshi Priyaa, S. Sathyapriya, L. Arockiam, “Nutrition Monitoring and Calorie Estimation Using Internet of Things (IoT)”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8, No. 11, pp. 2669-2672, 2019, DOI:10.35940/ijitee.K2072.0981119.
[19] Z. Zheng and K. Ruggiero, “Using machine learning to predict obesity in high school students”, In Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 2132-2138, 2017.
[20] K. Jindal, N. Baliyan, P.S. Rana, “Obesity Prediction Using Ensemble Machine Learning Approaches”, In Recent Findings in Intelligent Computing Techniques, Springer, pp. 355-362, 2018.

S.Sathyapriya

Research Scholar, St. Joseph's College

Data Analytics
Internet of Things

	All	Since 2017
Citations	3	3
h-index	1	1
i10-index	0	0

TITLE	CITED BY	YEAR
Survey on N-Queen Problem with Genetic Algorithm S Sathyapriya, R Stephen, VSJ Irudayaraj no 2, 54-59	3	2018
RULE EMBEDDED SEMANTIC ONTOLOGY BASED CLASSIFIERFOR IoTHEALTHCARE S Sathyapriya, L Arockiam Annals of the Romanian Society for Cell Biology, 10224-10231		2021
A Review Onsemantic Technology in Iot Healthcare S Sathyapriya, L Arockiam Annals of the Romanian Society for Cell Biology, 9810-9819		2021
Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India ADVK Dr. L. Arockiam, S. Sathyapriya, V.A. Jane International Journal of Recent Technology and Engineering (IJRTE) 8 (2 …		2019
Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning ADVK L. Arockiam, S. Sathyapriya, V.A. Jane International Journal of Recent Technology and Engineering (IJRTE) 8 (2 …		2019
PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR DIABETIC PREDICTION USING PIMA- INDIAN DATASET SS Dr. L. Arockiam, A. Dalvin Vinoth Kumar Journal of Emerging Technologies and Innovative Research (JETIR) 5 (12), 563-569		2018
Prevalence of Type-II Diabetics Association with PM 2.5 and PM 10 in Central Region of Tamil Nadu, India L Arockiam, S Sathyapriya, VA Jane, ADV Kumar		
Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning L Arockiam, S Sathyapriya, VA Jane, ADV Kumar		
IoT Based Architecture for Agricultural Farm Monitoring VA Jane, S Sathyapriya, L Arockiam TIRUCHIRAPPALLI-620 002, TAMIL NADU, INDIA		