DATA PREPROCESSING TECHNIQUES FOR IOT BASED IRRIGATION SYSTEM

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

V. A. JANE, MCA., MBA., M.Phil.,
(BDU Ref. No. 01050/Ph.D. K10/Computer Science/Full Time/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.

MAY 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.,*
Associate Professor in Computer Science
Director of MCA

CERTIFICATE

This is to certify that the thesis entitled **"DATA PREPROCESSING TECHNIQUES FOR IOT BASED IRRIGATION SYSTEM"** submitted by **Mr. V. A. JANE,** 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 him 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: **(L. Arockiam)**

Place: Tiruchirappalli *Research Supervisor*

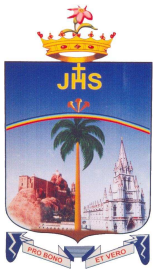
V. A. JANE,
Research Scholar
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 620 002.

DECLARATION

I hereby declare that the work embodied in this thesis entitled **"DATA PREPROCESSING TECHNIQUES FOR IOT BASED IRRIGATION SYSTEM"**, 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: **(V. A. JANE)**

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	**V. A. JANE**
2	Course of Study	Ph.D., Computer Science
3	Title of the Thesis/Dissertation	**"DATA PREPROCESSING TECHNIQUES FOR IOT BASED IRRIGATION SYSTEM"**
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	20.04.2022

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

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

Ouriginal

Document Information

Analyzed document	Jane 05-04-2022.docx (D134032243)
Submitted	2022-04-20T07:44:00.0000000
Submitted by	Dorairajan
Submitter email	manavaidorai@gmail.com
Similarity	0%
Analysis address	manavaidorai.stjct@analysis.ouriginal.com

Sources included in the report

W	URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1 Fetched: 2020-04-25T11:59:21.2830000	1
W	URL: http://www.columbia.edu/~st2839/final.html Fetched: 2021-03-12T09:41:51.6730000	3
W	URL: https://www.mdpi.com/1424-8220/18/9/3122/htm Fetched: 2020-01-10T21:16:25.7430000	1

இதனை இதனால் இவன்முடிக்கும் என்றாய்ந்து
அதனை அவன் கண்விடல்.

குறள் - 517

After having considered, "this man can accomplish this, by these means", let leave with him the discharge of that duty.

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 his 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. P. Calduwel Newton**, Assistant Professor, Department of Computer Science, Government Arts College, Tiruchirappalli-620 022 and **Dr. S. S. Manikandasaran**, Assistant Professor & Associate Director, Department of Computer Science, Adaikalamatha College, Thanjavur-613403., for their stimulating motivation and valuable ideas.

I owe a debt of gratitude to **Rev. Dr. M. Aockiasamy Xavier** S.J., Principal of 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**, Associate Professor (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.

I wish to express my special thanks to **Dr. G. Vaitheeswaran** and **X. J. Immanuel Kant Roch,** for providing valuable suggestion 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.

My heartfelt thanks to my Father **M. Venice Savariraj**, Mother **A. Francis Xavier Flora**, sisters **V. Bibiana** and **V. A. Barbara** for their support and blessings to complete my research journey.

I also place on record, my sense of gratitude to one and all who, directly or indirectly, have their helping hand in 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.

V. A. JANE

ABSTRACT

IoT technology is not used appropriately in Agriculture sector, especially in irrigation field. The proposed work collects irrigation data by using IoT sensors namely humidity, soil moisture, temperature, anemometer and rain sensor. Data is collected in mutual test IoT environment which is an IoT environment where the same set of sensors are used in different locations of the field to ensure data reliability. Mutual test IoT environment is designed to avoid missing values and outliers. But noise, sensor errors and irrelevant features are present. The processing time is increased and Classifier accuracy is decreased by the presence of noise, sensor errors and irrelevant features. Traditional IoT data analytics techniques are not suitable for mutual test environments. So there is a need for new preprocessing techniques.

Jo's architecture is proposed for the IoT data collected in mutual test environment. It includes various phases such as data collection phase, preprocessing phase and SVM classifier phase. The pre-processing phase combines the proposed three techniques such as Technique for Detection and Removal of Noise in IoT Data by using Central Tendency (DaRoN), Technique for hAndling seNsor errOrs in Smart irrigation system (TANOS) and Ensemble Filter Based Feature Selection for IoT Agriculture Data (MESIA). These techniques enhance classification accuracy of the proposed work.

Collected data have noise in the form of repetitive values, point noise, continuous noise, and class noise, attribute noise and collisional values which are not handled by the existing techniques. So, DaRoN technique is proposed to remove these noise. Generally, noise removal technique follows three stages of processing that is robust (detection of any analysis errors to make the data standardized), filtering (using

various measures to remove noise) and polishing (Replacing error values). Each stage requires separate techniques. But in this proposed work, it combines these three stages into a single step by using the timestamp value of sensors and central tendency measures. By using timestamp value, robust and filtering are done and after which polishing is done by using the central tendency measure.

Outliers do not directly exist in the proposed environment, but missing values are present in the form of sensor errors. Sensor error occurs when the sensor fails to collect data. There are various reasons behind sensor errors, such as connection failure, power failure and sensor failure. DaRoN technique removes the noise but ignores the sensor errors (missing values). Generally missing values are classified into three types. They are Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). Among these, MAR and MCAR are not harmful but MNAR is harmful which is handled in the proposed mutual test IoT environment. In order to improve classifier accuracy, neighbor value and neighbor mean values are used to replace the missing values. Not a Number (NaN) error exists in the proposed mutual test IoT environment which is removed by using the neighbor value. If the error is found in parent sensor then the mean value of neighbor is used for replacement. By doing this replacement, TANOS technique was proposed to remove sensor errors. After that, the dataset is fed to the SVM classifier to check the accuracy. Finally, the TANOS technique is compared with existing error removal techniques and yields high rate of accuracy than others.

MESIA Technique ensembles the univariate and multivariate filtering by using mean and threshold values as supporting factors. Initially, Multivariate filtering is performed in MESIA for which mean value is calculated and subsets are selected. After selecting the best subset, environment based threshold values are used to

perform univariate filtering to eliminate the irrelevant features. Positive (ρ) and negative (-ρ) correlations are used in univariate filter which eliminates irrelevant features based on environmental conditions. This process enhances the classifier accuracy and reduces the machine learning model building time. The proposed technique is compared with existing techniques based on the classification accuracy. Hence, the proposed technique proves that the accuracy of the classifier is increased and the training time are reduced.

PAPERS PUBLISHED

1. **V. A. Jane** and L. Arockiam, **"IoT Data Preprocessing - A Survey",** Webology, Vol. 18, No. 6, pp. 2070-2080, ISSN: 1735-188X, 2021.
2. **V. A. Jane** and L. Arockiam, **"CID: Central Tendency Based Noise Removal Technique for IoT Data",** Webology, Vol. 18, No. 6, pp. 2210-2217, ISSN: 1735-188X, 2021
3. Dr. L. Arockiam, S. Sathyapriya, **V. A. Jane** and 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, Vol.8, No. 2, pp. 6440-6444, ISSN: 2277-3878. (**Scopus Indexed**)
4. Dr. L. Arockiam, S. Sathyapriya, **V. A. Jane** and A. Dalvin Vinoth Kumar, "**Prevalence of Diabetes Mellitus in Tiruchirappalli District using Machine Learning**", International Journal of Recent Technology and Engineering, Vol.8, No. 2, pp. 6400-6403, ISSN: 2277-3878, (**Scopus Indexed**)
5. **V. A. Jane** and L. Arockiam, "**DaRoN: A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency**", Annals of the Romanian Society for Cell Biology, Vol. 25, No. 2, 2021, pp. 3197 – 3203, 2021, ISSN: 1583-6258. (**Scopus Indexed**)
6. **V. A. Jane** and L. Arockiam, "**Survey on IoT data preprocessing**", Turkish Journal of computer and mathematics Education, Vol.12 No. 9, pp. 238-244, 2021.(**Scopus Indexed**)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	-	Artificial Intelligence
ANN	-	Artificial Neural Network
ARM	-	Advanced RISC Reduced Instruction
BM3D	-	Block Matching and 3D Filtering
CEEDMAN	-	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
CFS	-	Correlation based feature selection
COAP	-	Constrained Application Protocol
COOJA	-	Contiki OS JAva simulator
CSV	-	Comma Separated Values
DaRoN	-	Detection and Removal of Noise in IoT Data by using Central Tendency
DFSS	-	Minimal Discriminating Feature Subset Selector
EF	-	Ensemble Filter
EM	-	Expectation Maximization
EWUSC	-	Error Weighted Uncorrelated Shrunken Centroid
FBCF	-	Fast Based Correlation Feature Selection
FCBF#	-	Fast Correlation Based Feature Selection #
FCFBiP	-	Fast Correlation Based Feature Selection in Pieces
FCFS	-	First Come First Served
FDIR	-	Fault Detection Isolation and Recovery
FKMI	-	Fuzzy K Means Clustering Imputation
FPM	-	Feet per Minute

HD	-	Hot Deck Imputation
IIoT	-	Industrial Internet of Things
Inf	-	error Infinity
INFFC	-	Iterative Class Noise Filter
IoT	-	Internet of Things
IPF	-	Iterative Partitioning Filter
ISTM	-	Incremental Space Time-based Model
KNN	-	K Nearest Neighbor
KNN	-	K Nearest Neighbor clustering
LAMSTAR	-	Large Memory Storage and Retrieval Neural Network
LOCF	-	Last Observation Carried Forward
LOOCSFS	-	Leave One Out Calculation Sequential Forward Selection
LSTM	-	Long Short Term Memory
MAPE	-	Mean Absolute Percentage Error
MAR	-	Missing at Random
MCAR	-	Missing Completely at Random
MCFS	-	Multi Cluster Feature Selection
MESIA	-	enseMble filtEr based feature Selection for IoT Agriculture Data
MLFSSM	-	Multilevel wrapper Feature Subset Selection Method
MNAR	-	Missing Not at Random
mRMR	-	Minimum Redundancy Maximum Relevance
MTL	-	Meta Learner
MTO	-	Multi Tracker Optimizer

NaN	-	Not a Number
NLP	-	Natural Language Processing
NOCB	-	Next Observation Carried Backward
PAoI	-	Peak Age of Information,
PCA	-	Principal Component Analysis
PD	-	Pairwise Deletion
PMF	-	Probabilistic Matrix Factorization
QoS	-	Quality of Service
RAM	-	Random Access Memory
RMSE	-	Root Mean Square Error
RNN	-	Recurrent Neural Network
SPARQL	-	SPARQL Protocol and RDF Query Language
SVM	-	Support Vector Machine
TANOS	-	Technique for hAndling seNsor errOrs in **S**mart irrigation system
WoT	-	Web of Things
WSN	-	Wireless Sensor Network
XGBoost-LSTM	-	Extreme Gradient Boosting Long Short Term Memory

Chapter – 1

Introduction

Chapter – 1
Introduction

IoT technology is not deployed properly in the agriculture sector, especially in the irrigation stage of a crop development. Traditional IoT data analytics techniques are not suitable for mutual test environments, which is adopted for avoiding missing values. Thus, there is a need for proposing new pre-processing techniques. The chapter 1 will be discussing various aspects of Internet of Things (IoT) with its wider applications and specific to Agricultural sector. Then, the data pre-processing will be briefed, following which basics and fundamentals of machine learning methodologies will be given. Then, the necessity of moving to the smart agriculture and its role are detailed.

Finally, the motivation behind the research, problems identified after knowing various limitations, scope of doing the research work and projected objectives of the aimed work are added.

1.1. INTERNET OF THINGS

Kevin Ashton defined “Internet of Things (IoT) as the network of physical objects or things embedded with electronics, software, sensors and network connectivity, which enables these objects to collect and exchange data” [**Ash, 09**]. Many sensing capabilities have been wide opened after the emergence deployment of Wireless Sensor Networks (WSN). Thus, many measuring and inferring needs have got proliferated and created Internet of Things [**Pat, 16**].

In all the sectors including agriculture, the count of interconnected devices get a hike for measuring / monitoring desired variables. Thus, the deployment of IoT in various fields have increased drastically.

Various functions of IoT could possibly be viewed as benefits. These are:

- Any category of devices could be interconnected without any constraints in its configuration, which vary from one device to another.
- Any user without any constraint on the count can monitor and maintain diversified data collected by all the connected devices.
- Any paths or networks could be used to facilitate the communication between each other.
- Neither the locations of the device nor user is a constraint, which enables easier accessibility of various connected devices.

1.1.1. Evolution of IoT

The IoT was extensively found deployed in the arena of wireless communication, where it evolved till the development of smart cities [**Can, 18**]. But, now a days, the IoT had been started using in the agricultural sector as well owing to the need of practically knowing every data involved in it. However, the skills and expertise required to keep-up with this evolution of IoT have been a question till now.

The evolution of IoT was realized in the following areas [**Can, 18**]:

- Connectivity
- Markets and enabling technologies
- Security

1.1.2. IoT Applications

As [**Kot, 18**] identified, IoT is deployed in various sectors including but not only limited to health care, education, envrionemntal investigation, automobile and agriculture, etc.

1.1.3. IoT-based Irrigation System

As the Agriculture is being regarded as the backbone of the Indian growth and development, the steps taken to improvise various processes in the agricultural crop cultivations have seen a hike. Especially, to facilitate smoother data transfer in the most important phase of irrigation monitoring, IoT-based Irrigation Systems are being proposed.

The need for monitoring systems in the Irrigation phase are as follows:

- Augmented Weed pressure
- Surface run-off
- Nitrogen leaching
- Unexpected yield losses
- Occurrence of diseases.

1.2. DATA PREPROCESSING

Any data which is being processed or transmitted can be improvised further well-advance of the actual process to provide better outcomes **[Alc, 19]**. This process is known as Data pre-processing.

There are many techniques adopted for pre-processing any given data, which are as follows:

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

1.2.1. Data Cleaning

IoT data will come from different sources of varying formats and structures, thus the need for the data to be pre-processed is higher. This process of processing the data to do the following functions are called Data Cleaning.

- To remove noisy data
- To handle sensor errors (Missing values)
- To select the best features from the collected data

1.2.2. Data Integration

According to [**Poi, 21**], Data Integration is the process of facilitating the availability of data in a single uniform view after collecting those data from varied sources. It helps in better accessibility of data in any application including but not limited to IoT. Having known its importance, the studies toward the Data Integration have started evolving from the works of [**Len, 02**], however identifying various challenges in achieving it has to be done comprehensively [**Sto, 18**].

1.2.3. Data Reduction

According to [**Ter, 21**], Data reduction is nothing but the process of reducing any volume of data to indicate it in a smaller volume. However, the integrity of the data should not be compromised.

Many works started using many kinds of data reduction techniques like Wavelet Transform, Attribute Subset Selection and Principal Component Analysis (PCA), etc. Some of those techniques were found including but not limited to [**Alq, 19**], [**Bev, 93**], [**Bev, 03**], [**Tod, 86**]. This reduction of data will be beneficial when dealing IoT data too.

1.2.4. Data Transformation

According to [**upG, 21**], Data transformation is a method merging the unstructured and structured data for the sake of investigating it later. This process was especially useful when dealing with data in the cloud environment.

Improvising the data in the pre-processing stage itself was found to be efficient as per many methods including but not limited to [**Cal, 17**], [**Zho, 19**] When dealing with data from varied sources, these methodologies will be beneficial.

1.3. MACHINE LEARNING

Machine learning is a subset of Computer Science and Artificial Intelligence (AI), that concentrates to achieve the human learnability by the usage of various algorithms and data [**IBM, 21**]. These methodologies were found applied in many applications like prediction and classification, etc.

1.3.1. Basic Concepts

As [**Goo, 16**] discussed, the Machine learning methods could be categorized into the following types based on the experience that it gains via various tasks:

- Supervised learning
- Unsupervised learning

- Multi-instance learning
- Semi-Supervised learning
- Reinforcement learning

Various tasks would be defined in any machine learning method [**Goo, 16**] and a few among them are listed below.

- Classification
- Regression
- Machine Translation
- Anomaly detection
- Imputation of missing values
- Synthesis and sampling
- Denoising and many more

1.3.2. Machine Learning for IoT Analytics

With much efforts taken to improve the ways of applying IoT, the deployment of machine learning concepts have started becoming prevalent in the research area. [**Cui, 18**] made a comprehensive study of the Machine Learning for IoT to identify various issues pertaining to it. Some of the challenges identified [**Cui, 18**] were:

- Unstructured data
- Constraint of computer resources
- Practical analysis of data on the net

Thus, the work started to focus on the practical analysis of the data as found in the literature like: [**Adi, 20**], [**Dab, 19**].

1.3.3. Machine learning SVM model

One such method of machine learning known as Support Vector Machine (SVM) was proposed for the purposes of classification and regression [**Gee, 21**]. This is a supervised learning methodology, which classifies the data points.

1.4. Smart Agriculture

Making the traditional agricultural methodology into more intelligent systems has become the need of the hour in order to tackle various difficulties in the developmental stages of a crop. For instance, monitoring and maintaining the irrigation stage has become a necessary process in order to overcome the issues of excess water supply and abnormal yield losses.

1.4.1. Role of IoT in Smart agriculture

Agricultural Monitoring using IoT has eased the conventional way of agricultural monitoring, wherein the practical data can't be handled. Thus, diversified quantity of data shared between the interconnected devices could be efficiently handled with the utilization of IoT to securely transmit the data to the recipient with less loss and increased accuracy. Also, an environmental friendly agricultural processes were only possible with the application of IoT [**Pat, 19**], [**Sha, 16**].

1.5. MOTIVATION

Since the quality of data after getting transmitted might be degraded or less effective, it becomes necessary to pre-process the data. In particular, when dealing with IoT data, various devices might be involved, owing to which the handling of heterogeneous data can't be avoided. Hence, pre-processing of the data becomes significant in IoT applications.

Furthermore, any method of classification or predicting might not perform well when there is unnecessary noise and errors in the input given to it even before the feature selection. Thus, it also becomes important to avoid such instances and improvised the input to the feature selector to perform classification or predicting tasks in a desirable way.

1.6. PROBLEM DEFINITION

IoT is a predominant technology which makes many applications smarter using its features. In the past, gathering data in agriculture environment was a difficult task especially in irrigation monitoring systems but IoT removes all those strenuous part with the help of sensors. Here, sensors play a vital role in data collection and generates enormous data every day. These data contain missing values, noise, outliers, and duplicate values. If any one of the above issues present in the collected data, then it will reduce the quality of outputs. Data cleaning is one of the important tasks in data preprocessing. There are some existing techniques for data cleaning, but the proposed environment used mutual test to collect data. So, existing techniques are not suitable for handling the collected data. There is a need for new preprocessing techniques to remove noisy data, sensor errors and irrelevant features to attain high accuracy rate. With the deployment of the IoT in agricultural sector being beneficial mostly, there are still some areas of concern, which needs attention. Those concerns are listed below.

- A dynamically changing IoT environment could inevitably develop noisy data and sensor errors, which left and attended before processing into the proposed irrigation monitoring system might leave us with more inaccuracies. Thus, the

assessment of these sensor errors (missing values) and removal of noisy data becomes an imperative pre-processing step.

- Eliminating the missing data and ignoring noisy data lead to incorrect analytical results. So, proper handling of the missed data and noisy outputs from the sensor is necessary to come up with nearly close results as desired.
- Furthermore, incorrect accuracies in the features selection cause the Model building time to increase, which thereby decrease the accuracy of the model.

1.7. SCOPE AND OBJECTIVES

The scope and objectives of the proposed work will be discussed below.

1.7.1. Scope

IoT technology is not deployed properly in the agriculture sector, especially in the irrigation stage of a crop development. Traditional IoT data analytics techniques are not suitable for mutual test environments, which is adopted for avoiding missing values. Thus, there is a need for proposing new pre-processing techniques.

Furthermore, processing time was found increasing and classifier accuracy was found decreasing due to the presence of noise, sensor errors, and irrelevant features being selected. As a result, the scope of the research gets narrowed down to propose novel pre-processing techniques for redressing all the reported issues.

1.7.2. Aim and Objectives

The aim of the proposed research work is to propose a IoT data pre-processing architecture to clean the noisy data, handle sensor errors efficiently, and to favorably

select the best possible features to enhance the accuracy of IoT data analytics. The aim could be achieved by the following projected objectives:

- To propose a Technique that can remove the noise in IoT data for improving the classifier accuracy.
- To propose a Technique for removing all the identified sensor errors in IoT data towards the yielding of higher accuracy rate.
- To propose a Technique for selecting the best features in the collected dataset towards the efficient pre-processing of the IoT data.

1.8. ORGANIZATION OF THE THESIS

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

Chapter 1 gives the introduction of IoT along with its evolution and applications with reference to IoT-based irrigation system. Then, the pre-processing techniques are briefed with its various types. Basics of machine learning method is discussed and overview of IoT Analytics and SVM models are given. Smart agriculture and the role of IoT are then briefed. Finally, the research motivation, problem definition, scope and objectives are presented one by one.

Chapter 2 discusses various existing methodologies and areas of improvement in it with reference to IoT Data Preprocessing Techniques after summarizing the Data Mining and Machine learning techniques.

Chapter 3 explains the proposed Jo's architecture of pre-processing the data with the combined efforts of DaRoN, TANOS, and MESIA techniques along with its

background, literature history, need, and objectives, etc. The effectiveness of the proposed methodology will then be validated, interpreted and summarized.

Chapter 4 explains the proposed DaRoN technique to detect and remove noise in the collected irrigation data by using Central Tendency. After discussing its background, literature history, need and objectives, etc. The effectiveness of this proposed methodology will then be validated, interpreted, and summarized.

Chapter 5 explains the proposed TANOS technique to handle the sensor errors in the collected irrigation data by using neighbor values. The background, literature history, need and objectives, etc. pertaining to the proposed work are also given. Finally, the validation, interpretation, and summarization of the proposed methodology will be done.

Chapter 6 depicts the devised MESIA technique to select features by using the ensemble filter. Various aspects like background, objectives, literature history, and need of the work are conducted, after which the results are presented.

Chapter 7 summarizes the features of the proposed Jo's architecture for preprocessing IoT irrigation data by using the three techniques (DaRoN, TANOS, and MESIA) and its findings and limitations are presented. Finally, the recommendations for the future work are given.

Chapter – 2

Review of Literature

Chapter – 2

Review of Literature

2.1. INTRODUCTION

With the immense attention gained towards the IoT systems in various applications including Agriculture, Healthcare, Education and many more, challenges and limitations in applying those IoT systems keeps on increasing. Also, IoT itself being an evolving topic, many researchers have started to take up various methods in improvising it in some way or other. One such way was data pre-processing, which could sort out the problem of data quality encountered in the IoT system.

2.2. BASIC CONCEPTS AND DEFINITIONS

The basic concepts and definitions pertaining to IoT and pre-processing techniques will be discussed below.

2.2.1. Internet of Things

Kevin Ashton defined "Internet of Things (IoT) as the network of physical objects or things embedded with electronics, software, sensors and network connectivity, which enables these objects to collect and exchange data" [**Ash,09**]. IoT data collection is the process of using sensors to track the conditions of physical things. Devices and technology connected over the Internet of Things (IoT) could monitor and measure data in real time. The data would be transmitted, stored, and be retrieved at any time depending upon the need.

2.2.2. IoT Data Preprocessing

According to **[Jan, 21a]**, Data preprocessing is nothing but the process of converting any raw data into a more comprehendible format. This step of pre-processing

is also an indispensable step in the data mining since working with raw data is tedious or impossible. Thus, the data quality should be verified before implementing any data mining or machine learning methods. This quality of data could be verified by the following variables:

- Timeliness
- Completeness
- Interpretability
- Accuracy
- Believability
- Consistency

Not only the implementation of the machine learning or data mining methods require data pre-processing, but also the IoT system integrated with the above-mentioned methods require data pre-processing, as big quantity of data is being handled. The chances of uncertainty when dealing with huge data might increase in the IoT systems without data pre-processing [**San, 18**]. Some of the issues which IoT data pre-processing aims to redress will be discussed in the following sections.

(i) IoT data noise handling

Noise in the collected dataset refers to meaningless information such as distorted values, repeated values, error values, and null values, etc., **[Jan, 21a]**. As various industries have already started deploying IoT-based systems, one of the important issues like predictive maintenance has become indispensable, which would be only possible by the elimination of noises Liu et al., [**Liu, 20a**]. Without the noise removal, anomaly detection would also be impossible in data mining and machine learning methods **[Liu, 20b]**. Thus, IoT implementation would also be limited.

For instance, a typical noise eliminating process was adopted by Sulthana et al., [**Sul, 18**] to improvise the heart monitoring system. The work made use of modified Circular Leaky Least Mean Square (CLLMS) method to tackle the noise issue. Hence, they were able to tackle the noise issue to avoid the humiliation of the heart-based signals and higher amplitude rates. Table 2.1 shows the summary of noise handling papers.

Table 2.1: Review on Survey Papers published in Noise Handling

Citation	Techniques /Algorithms /Methods/ Key words	Work	Domain
[**Hir, 15**]	Dimensionality reduction, Markov Blanket Filtering, Information Gain Ranking, Unconditional Mixture Modelling, Attribute noise, Class noise, Error-Weighted Uncorrelated Shrunken Centroid (EWUSC), Minimum Redundancy Maximum Relevance (mRMR), Correlation-based feature selection (CFS), Gradient-based-leave-one-out gene selection, Leave-one-out calculation sequential forward selection (LOOCSFS), Gene Ontology	Irrelevant and redundant features removed using dimensionality reduction techniques.	High Dimensional Micro Array Data
[**Nat, 16**]	Mean, Standard Deviation and Skewness, Complementary Filter and Kalman Filter, Support Vector Machine (SVM)	Proposed filter based sensor fusion System to monitor user activities and remove noisy data.	Acceleration sensors, position sensors, vision sensors, audio sensors, temperature sensors and direction sensors data

[**Ram, 17**]	Concept drift detectors, Sliding windows, Online learners, and Ensemble learners	Survey on preprocessing and data reduction techniques, noise handling techniques and empirical analyses on existing methods	Streaming Data
[**Sud,18**]	Noise Types 1. Gaussian Noise 2. Poisson Noise 3. Salt and Pepper Noise 4. Speckle Noise Noise Filter Types 1. Linear Filter 2. Min Filter 3. Max Filter 4. Median Filter 5. Wiener Filter 6. Gaussian Filter 7. Guided Filter 8. Block Matching and 3D Filtering (BM3D) 9. Adaptive Fuzzy Switching Median Filter	Summarized the noises in digital image processing and Compared the performances of all noise filter used in digital image processing	Digital image processing
[**Wei, 18**]	Kalman Filter, Z-scoring and moving average filter	Study on Kalman filter performances in classification of noisein the chemical sensor data.	Chemical Sensor Data
[**Evg, 18**]	Last Observation Carried Forward (LOCF) Next Observation Carried Backward (NOCB), interpolation (linear, polynomial, Stineman) and moving average (simple, weighted, exponential), e Structural Model & Kalman Smoothing, ARIMA State Space Representation, Root Mean Square Error (RMSE) and	Study on sensor data preprocessing including noise handling and related techniques are discussed.	Streaming sensor data

	Mean Absolute Percentage Error (MAPE)		
[**Gup, 19**]	Filtering, Noise handling mechanisms	Reviewed noisy data handling methods and evaluated all the techniques and methods used to handle noise.	General Data Analytics
[**Cha, 19**]	Peak Age of Information (PAoI), First Come First Served (FCFS)	Reducing transmission time and increasing processing speed in IoT environment by using noise handling Techniques	IoT Data
[**Mor, 19**]	IoT, Sensors, Data Collection, Smart Applications	Surveyed all papers published related to IoT noise handling since 2015	IoT Sensors Data
[**Teh, 20**]	Sensor data quality(Fault Detection, Isolation, and Recovery (FDIR)), Sensor data error detection (Principal Component Analysis (PCA) and Artificial Neural Network (ANN)), Sensor data error correction (Association Rule Mining, K-Nearest Neighbor (KNN) clustering, tensor-based singular value decomposition, and Probabilistic Matrix Factorization (PMF))	Detailed Physical sensor data collection error and error detection and correction mechanisms.	IoT Data

(ii) Sensor errors (Missing values) handling

Owing to the increase in the remote processes in many applications like agricultural and environmental monitoring, wider variety of sensors have been deployed in all machine-machine communications. As the count of the sensors increase, the risks of errors steps in. These can't be avoided in the IoT data handling as well, where in the missing value was one among the issues reported when dealing

with diverse sensors [**Pen, 19**]. Not only the data handling gets affected because of missing value errors in the sensors, but also the reliability of any monitoring systems gets affected [**Liu, 20b].** Table 2.2 shows the summary of missing value (sensor error) handling papers.

Table 2.2: Survey papers on missing values handling techniques

Citation	Type	Algorithm, Technique, Keyword, Methods	Area	Work
[**Swa, 16**]	Survey	**Missing Data Ignoring Techniques** (List wise Deletion (Or Complete Case Analysis), Pairwise Deletion (PD)), **Missing Data Imputation Techniques**(Mean Value Imputation Method, Hot Deck Imputation(HD):, K-Nearest Neighbor Imputation (KNN):, K-Means Clustering Method:, Fuzzy K-Means Clustering Imputation (FKMI):, Regression Imputation:, Multiple Imputations), **Missing Data Model-Based Techniques**(Maximum Likelihood, Expectation-Maximization (EM) Algorithm	Data Mining	Classified missing data handling techniques.
[**Swe, 17**]	Survey	k-Nearest Neighbor, Privacy Protection 1.Heuristic –Based Techniques 2. Border Approach 3. Exact Approach 4. Reconstruction based association Rule 5. Cryptography based Techniques 6. Hybrid technique approach,	Big data	Surveyed on Data imputation and privacy

		Heuristic- based techniques(Data distortion method (Uniformly distributed noise, Normally distributed noise)Data blocking method)		
[**Kwa, 17**]	Review	Missing Values, Outliers, Trimming, Winsorization, Robust estimation method, Imputation analysis, Available case analysis, Complete case analysis	Statistical Data	Review on missing values and outliers handling
[**Pap, 18**]	Case Study	Single imputation, Multiple imputation MCAR, MAR, MNAR	Clinical Research	Classified single imputation and multiple imputation techniques.

(iii) Feature selection

According to Gonzalez-Vidal et al., [**Gon, 19**], feature selection is a method of determining the best possible features among the various data gathered in any applications like IoT, Machine learning, and Data mining, etc. Further, the author [**Gon, 19**] made use of a time-dependent energy effective feature selecting methodology for reducing the tedious task of selecting the best features towards the deployment of successful IoT-based smart city project. Likewise, many similar applications like agriculture and health care make use of it as and when needed.

2.3. IoT DATA PREPROCESSING TECHNIQUES

Whenever the data mining in an IoT application has to be improved, it is only possible with various data pre-processing techniques right from the anomaly identification to the repetitiveness detection in the data handled [**Zho, 19**].

Variety of application/ system that need the data to be processed in its initial stage include: intelligent city, intelligent transport, intelligent agricultural monitoring, intelligent medical care, intelligent building, and intelligent environment [**Kri, 20**].

2.3.1. Data Mining Techniques

Data Mining is the process of acquiring the information from a considered dataset to recognize the patterns, useful data and trends in it **[Age, 21]**. Without the initial data quality checks being carried out, the data mining process might be less effective irrespective of its range of applications.

Also, the data mining techniques would get categorized into the following types:

- Clustering
- Association
- Prediction
- Classification
- Sequential patterns

According to Savaliya et al., [**Sav, 18**], there were two variants available for data mining system, which are as follows:

- Distributed data mining system- The data are processed by transferring to a distributed node.
- Multi-Layered data mining system- The system got divided into four layers, namely: data administration layer, information gathering layer, event processing layer and data mining service layer.

2.3.2. Machine-Learning Techniques

Machine learning methods were deployed for the sake of gaining some insights from any piece of data **[Elb, 21]**. Also, most known topics revolving around the machine learning techniques include:

- Transfer Learning
- Clustering
- Ensemble Methods
- Regression
- Neural Nets/ Deep Learning
- Reinforcement Learning
- Word Embeddings
- Natural Language Processing (NLP)
- Dimensionality Reduction
- Classification

The most commonly used technique of machine learning was Support Vector Machine (SVM). However, it had become thing of the past because of many advanced methods adopted in machine learning.

For instance, A multiple processes aware methodology was proposed by Wang et al., [**Wan, 20**] for successfully predicting the speed of the wind. Data pre-processing were done initially to raise the reliability of the prediction mode. Likewise, Kumar et al., [**Kum, 20**] solved the transportation system issue of raising the Quality of Service (QoS) by deploying the innovative heuristic simulation optimizing process.

2.4. ANALYTICAL SURVEY OF EXISTING WORKS

Any literature review with an analytical survey won't be able to investigate how any response variable could be related to any particular explanatory variable or variables. For instance, plant diseases identification or managing the irrigation system in an agricultural-oriented application that can't be beneficial unless knowing various

factors/ variables involved it. Thus, here is an analytical survey pertaining to the various methodologies revolving around IoT and Data Pre-processing.

Dachyar et al., [**Dac, 19**] amalgamated the IoT articles published in the period between 2006 and 2018. The author addressed the most influential industries of IoT applications and listed the tools and methods used for handling such applications. Thus, they categorized the IoT application problems into two types, namely: before implementing IoT and after implementing IoT.

Kumar et al., [**Kum, 19**] delineated the IoT architecture to express the related technologies associated with each layer. The author elucidated the major key issues of IoT, namely: Security and privacy issues, Interoperability/ Standard issues, Ethics, law, regulatory rights, Quality of Service (QoS), Scalability, availability, and reliability. The relationship between IoT applications and big data analytics was explicated distinctly.

Andersen et al., [**And, 20**] delineated the relation between IoT and Big Data analytics by discussing the research summary of IoT data analytics. The author divided the IoT data analytics into two parts, namely: IoT for data collection and Big Data analytics techniques for processing that collected IoT data.

Su et al., [**Su, 19**] proposed a feature selection method to select features in correlation changing IoT environment. In this method, correlated features were clustered to monitor the changes. If any changes were found in the features, then the feature was moved to anomaly detection. Otherwise, Multi-Cluster Feature Selection (MCFS) was used for feature selection. This method reduced the false-negative value and improved the performance by 30 %, but this method is not applicable to the large data set.

Vandana et al., [**Van, 21**] presented a Minimal Discriminating Feature Subset Selector (DFSS) approach to select the best features in the IoT environment. This approach consisted of two parts, namely: feature ranking and selecting feature subset based on that rank. COOJA simulator with Constrained Application Protocol (COAP) was used for implementing the proposed approach. This approach is not applicable when the features have any relation and dependency between each other.

Mao et al., [**Mao, 19**] proposed a Multilevel wrapper Feature Subset Selection Method (MLFSSM) to select features in a medical data. This method handled the complex interaction problem by using features weight in each layer. These weights were revised from layer to layer. The topmost weighted layer was used for selecting the feature subset. The author used SVM based fivefold validation for 20 times to remove the low fitting problem, but there existed an overfitting problem. The multilayer processes used in this technique reduce the system speed and increase the algorithm complexity.

Egea et al., [**Ege, 17**] proposed a smarter IoT-based classifying method for quicker correlation feature selection by keeping in mind the environment of industrial perspective. They divided the feature space into many equal-sized fragment and prioritized the traffic data.

Radhakrishnan et al., [**Rad, 21**] proposed a Deep-RNN (Recurrent Neural Network) method on the data of Advanced RISC- Reduced Instruction Set Computing Machines, also known as ARM. The deep-RNN method took care of the Long Short-Term Memory (LSTM) to enhance the data processing. This method had three phases in it; In phase one, the ARM data was extracted as OpCodes; after that, data was converted to vector form; and in the final phase, best subset data was collected based

on the vector points. The endorsed drop-out strategy was helpful in removing the over fitting problem only on small data sets, but if the data set size is increased, then this method suffers from the over fitting problem. Even though the performance was increased by the Deep-RNN method, the execution time was higher.

Mohtashami et al., [**Moh, 19**] devised a hybrid filter-oriented feature selecting methodology to optimize the classification process of microarray datasets. They achieved this improvised feature selection by the deployment of the concepts of hesitant fuzzy as well as the rough sets. Repeated features were also removed during the feature selection method.

Gopika et al., [**Gop, 18**] juxtaposed the performance of machine learning based dimensionality reduction algorithms, namely Fast Correlation Based Feature Selection (FCBF), Fast Correlation Based Feature Selection # (FCBF#) and Fast Correlation Based Feature Selection in Pieces (FCFBiP) on IoT data. After dimensionality reduction, Correlation-based feature selection (CFS) was used to identify the discrete and continuous features. FCFBiP algorithm outperformed FBCF and FBCF# because features were split into pieces. These pieces make easier the subset selection process. If the features have a negative correlation, this algorithm is not applicable and work execution time is increased while working with FBCF algorithm.

Sundararajan et al., [**Sun, 20a**] made an effort for identifying the Sarcasm in the Twitter portal by using the multiple ruled-based ensemble feature selecting methodology. Since the sarcasm don't always reveal the actual meaning behind whenever it was used in social media like Twitter, the identification of it become tedious. However, the identification of the Sarcasm was made easier with the investigation of emotional state of people.

A semantic dependent Interoperability methodology was suggested by [**Cim, 20**] for the Web of Things (WoT) by keeping in mind the Heterogeneous type of IoT Ecosystem. The work presented a SPARQL (SPARQL Protocol and RDF Query Language) query-oriented phenomenon for creating transparency intermediary to every IoT devices, that published heterogeneous type of data.

Successful traffic forecasting in the base station was done by Du et al., [**Du, 19**] with the deployment of XGBoost-LSTM and enrichment of the features. Pre-processing towards the recovery of the missing values were done was an initial step and then mining of the tidal characteristic was done with the feature engineering.

Lin et al., [**Lin, 19**] proposed an IoT-based malfunction identification and calibrating phenomenon-based solution known as the Sensor Talk. The aging sensors were successfully identified by this proposed methodology in order to prevent it from any potential malfunction. They devised both the simulation and analytical models to identify the malfunction well in advance by choosing the identification delay.

Medapati et al., [**Med, 20**] proposed an improvised adaboost-based Large Memory Storage and Retrieval Neural Network (LAMSTAR) in the application of IoT to successfully recognize the face for enabling the intelligent cities. This work was proposed to tackle the safety related concerns that could take place in any city. An algorithm of Perona-Malik diffusion was applied initially to the IoT device captured images and then the geometric model was created for that image to extract the features of the face effectively with the help of Fisher linear discriminant investigation.

Farahani et al., [**Fah, 20**] developed a collaborative intelligent machine learning system for medical care application by the distribution of intelligence across the layers of device, cloud, and fog/ edge. This system aided any medical professionals to

continuously track various data subjects irrespective of time and place with practical insights.

A wearable sensor dependent IoT methodology was devised by [**Hui, 20**] to aid the sportsmen in tracking various data pertaining to their health so that early curing of the sportsmen could be achieved. Since the early curing was facilitated by this method, the players who have faced a medium to critical survey too could return to the playing phase from their rest phase.

IoT is a predominant technology which makes many applications smarter using its benefits. In the past, gathering data in various applications including agriculture environment was a difficult task especially in irrigation monitoring systems, but IoT had started removing all those strenuous part with the help of sensors. However, data of the sensors contain missing values, noise, outliers and duplicate values. If any one of the above issues is present in the collected data, then it will reduce the quality of outputs. Thus, the future methods should include irrigation data, by using IoT sensors, namely: humidity, soil moisture, temperature, anemometer and rain sensor. Also, data should be collected in mutual test environment to avoid missing values and outliers.

2.5. ISSUES AND CHALLENGES

There are many issues in implementing and using IoT-based systems. This happens not only in one specific area, but happens with any applications like agriculture, smart city, traffic, and health care, etc. Some of the issues identified in the earlier literature are as follows:

- Interoperability issues were compressively reviewed by [**Nou, 19**] to show its importance in IoT.

- Sinha et al., [**Sin, 22**] through their comprehensive investigation identified many issues like standardization of IoT, Regulatory problems, Data unreliability, and several market issues, etc.
- Various challenges in intelligent farming IoT application specific protocols were summarized by [**Gla, 20**]
- Farooq et al., [**Far, 20b**] reviewed various challenges like security issues, lack of expertise in technologies, expensive cost, Unreliability, scalability issues and interoperability issues, etc.
- Nizetic et al., [**Niz, 20**] made a detailed review of challenges and scope for improvements in various IoT-based applications including intelligent city, energy maintenance, Food domain, Asset management, and waste management, etc.
- Issues in a finance aware system for Io-based implementation were discussed by Ruan et al., [**Rua, 19**]
- Issues pertaining to the farming in arable lands were discussed by Villa-Henriksen et al., [**Vil, 20**]
- Sensor errors/ outliers/ missing values were studied and interpreted by works of Li et al., [**Li, 20**], Teh et al., [**Teh, 20**], Guillen-Navarro et al., [**Gui, 21**], Tkachenko et al., [**Tka, 21**] and

Thus, the future researchers have to aim for proposing an IoT data pre-processing architecture to clean the noisy data; to handle sensor errors; and to select the best features to enhance the accuracy of IoT data analytics.

2.6. CHAPTER SUMMARY

This chapter had discussed various works found in the literature pertaining to the IoT, data pre-processing techniques, machine learning techniques and various challenges along with the briefing of the involved basic concepts was given. A detailed analytical survey was also done to realize the phenomenon on which they were proposed and implemented.

Chapter – 3

Jo’s Architecture for IoT Data Preprocessing

Chapter – 3

Jo's Architecture for IoT Data Preprocessing

3.1. INTRODUCTION

IoT data collection is the process of using sensors to track the conditions of physical things. Devices are connected over the Internet and it can monitor and measure data in real time [**Kun, 21**], [**Bor, 14**]. The data are transmitted, stored, and can be retrieved at any time. There exists a blind spot problem in traditional IoT applications [**Far, 20a**], which is shown in figure 3.1. Figure 3.1 shows the methods of collecting data from different locations and the problems which arise in it. Every sensor has a limited sensitivity range. Data collected beyond the sensitivity range leads to the malfunction of IoT applications. For example, in smart agriculture, the collected data are applied to the entire agricultural field. In some cases, data are not common in all fields. Lets consider, If a temperature sensor used in an irrigation field shows the value 20 °C if it is applied to the entire field, but the actual value in the blind spot area (non-sensitivity area) is 40 °C. Then this data is not reliable. If an IoT application uses these kinds of data, the application is not suitable for real-time. The low-quality sensor is also one of the reasons for the blind spot problem.

Data are collected under a mutual test environment [**Yib, 19**] to avoid the data reliability (blind spot) problem and to make the application efficient in real-time. The mutual test environment is an IoT environment where the same set of sensors are used in different locations of the field to ensure data reliability.

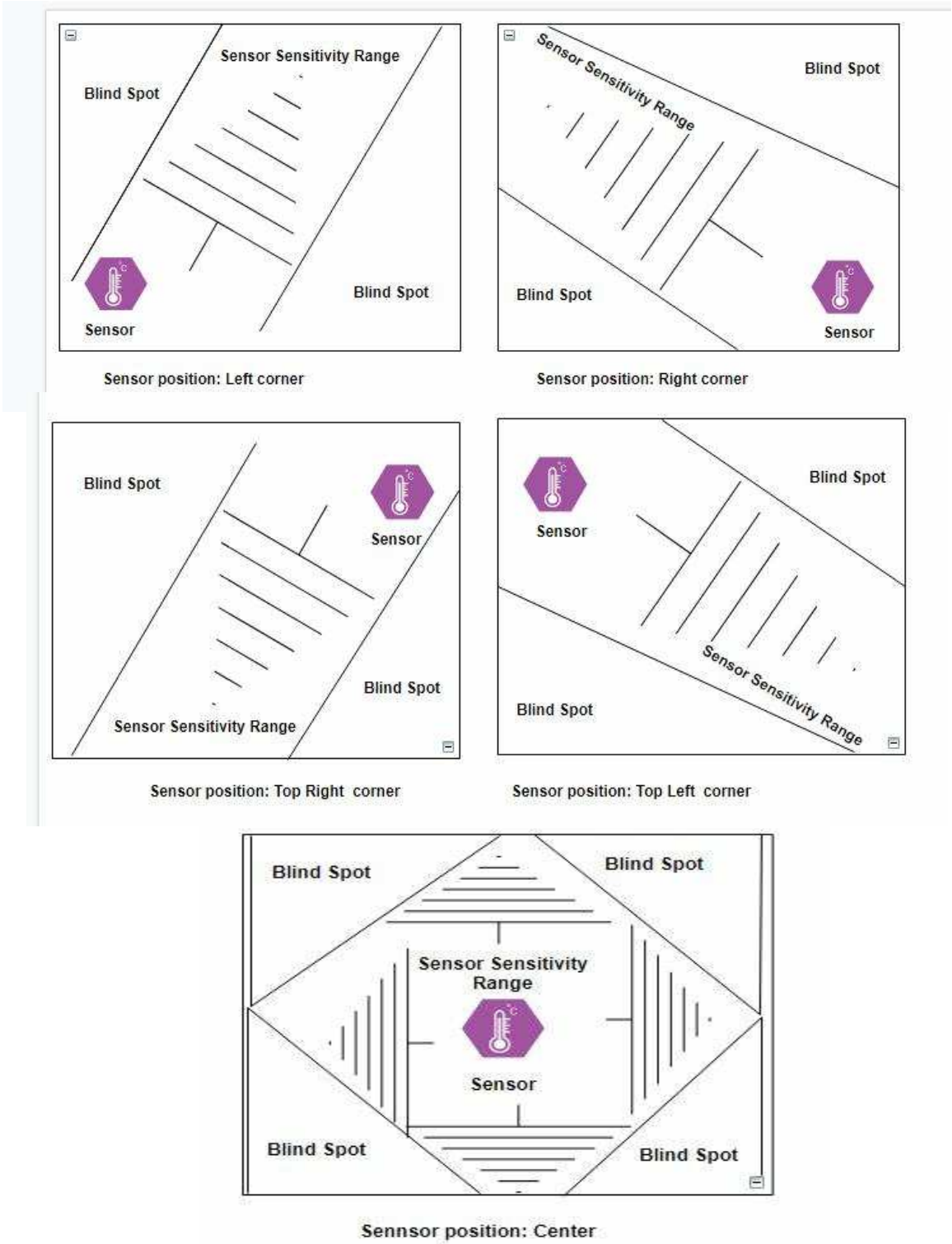
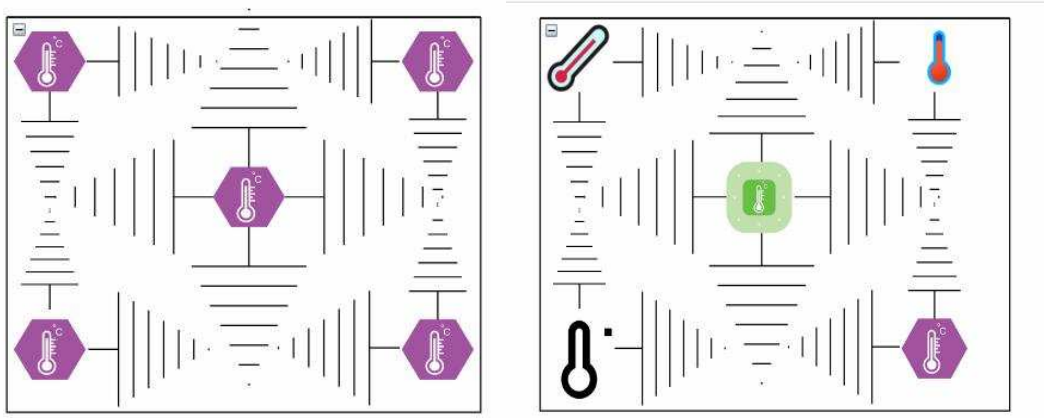


Figure 3.1: Data collection in different locations

In a multi-test IoT environment, it uses different types of sensors. They are placed in different places of a field for the same purpose. Figure 3.2 shows the example of the data collection under mutual and multi-test environments.



Mutual Test Environment **Multi-Test Environment**

Figure 3.2: Mutual Test and Multi Test

Table 3.1: Comparison IoT data collection methods

Type of IoT Environment	Traditional	Mutual Test	Multi-test
Data Quality	Low	Medium	High
Classifier accuracy	Normal	High	Very high
Cost	Less	Average	Expensive
Problem	Huge noise, missing values and outliers	Less noise, Partially eliminates missing values and outliers are eliminated.	Sensor collision problem. Parent sensor identification is difficult
Techniques	Traditional data mining techniques	Customized Data mining technique	Customized Data mining technique
Data Size	Small	Large	Large
Processing	Near real time	Real time	Real time

In a mutual test IoT environment, the same type of temperature sensor is used in different places of the field, whereas in a multi-test environment different types of temperature sensors are used in different places of the field. Comparison between IoT data collection methods is discussed in Table 3.1.

The mutual test environment is selected at a low cost while comparing with the multi-test environment. In this proposed work, Data collection is based on mutual test environment. That is the same set of sensors are used in the same field for data collection [**Thi, 21**]. This partially eliminates the missing values and the outlier problem in the IoT environment. IoT technology removes the barriers in traditional data collection [**Liu, 21a**]. But, IoT data is collected from different sources with varying formats and structures. A dynamically changing IoT environment inevitably develops noisy data and sensor errors problems [**Ana, 21**]. Traditional IoT data collection problems, namely missing values and outliers are avoided in a mutual test environment. The collected mutual test IoT data have noise in the form of repetitive values, point noise, continuous noise, attribute noise, class noise and collisional values, and also Not a Number (NaN) sensor data error exists in the collected data [**Wat, 21**]. The estimation of sensor errors (missing values) and removal of noisy data has become an imperative pre-processing step [**Jan, 21b**]. The collected data need to be pre-processed to remove noisy data, to handle sensor errors (Missing values) and to select the best features from the collected data. Eliminating the missing data and ignoring noisy data lead to erroneous analytical results. Model building time is increased and accuracy is decreased by irrelevant features.

3.2. BACKGROUND STUDY

3.2.1. Pre-processing techniques

Preprocessing is an important phase, in data analytics [**Afs, 21**]. The quality of the decision-making system directly relies on the preprocessing process. Many data science researchers addressed that, Preprocessing is the least enjoyable part of their research [**Gil, 16**]. There are four steps in data preprocessing namely data cleaning, data reduction, data integration, and data transformation. Data cleaning is the process of making the collected data error and noise-free [**Rod, 21**]. There are four steps in data preprocessing namely data cleaning, data reduction, data integration, and data transformation. Data cleaning is the process of making the collected data error and noise-free. Data integration is the process of merging data from heterogeneous sources [**Ahm, 22**]. Data reduction is the process of optimizing the amount of storage consumed [**Abd, 21**]. Data transformation is the process of converting the data into a common format [**Yu, 21**]. This research is done on data cleaning. There are two steps in data cleaning namely handling missing values and noisy data removal. Traditional preprocessing techniques are not capable of handling mutual test data. So, these methods have to be customized to handle mutual test data. Preprocessing steps are shown in figure 3.3. The highlighted techniques are carried out in the proposed research work.

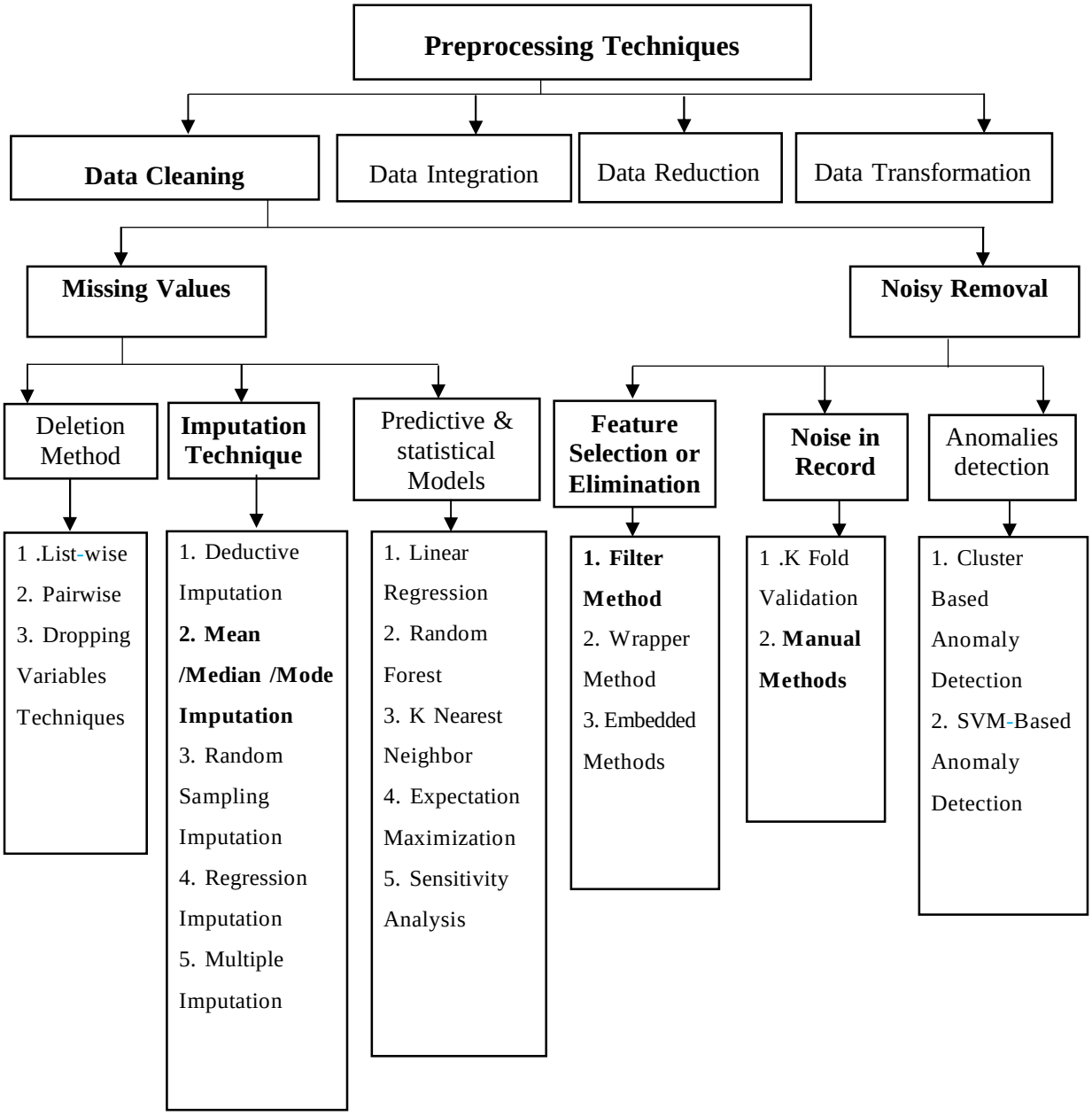


Figure 3.3: Preprocessing Techniques

3.3. RELATED WORK

Yi et al., [Yi, 19] recommended a solution called Sensor Talk to automatically detect potential sensor failures and calibrate the aging sensors semi-automatically. The author suggested two new data collection methods, namely, mutual test and multi test to avoid difficulties in data collection. A mutual test environment uses the same set of sensors in different locations whereas, a multi-test environment uses different set of sensors to collect data. Sensor Talk used multiple mutual tests to identify the failed sensor and actuator. Error detection delay was increased while reading sensors and actuators at the same time. Sensor Talk achieved a 0.7% false detection probability of sensor errors.

3.4. NEED FOR THE RESEARCH

IoT technology is facing more number of hurdles in the Agriculture sector, especially in the irrigation sector. Because of the data quality, many IoT based agricultural applications fail in real time. The proposed Jo's architecture collects irrigation data, by using IoT sensors namely humidity, soil moisture, temperature, anemometer and rain sensor. Data is collected in the mutual test environment to avoid missing values and outliers. But, there exists noise, sensor errors and irrelevant features. The table 3.2 shows the types of noises that exist in the collected data. For example, while comparing mutual test IoT environment and traditional IoT environment, repetitive values are not a big threat to traditional IoT applications, but repetitive values are a big threat in the mutual test environment. The reason behind this noise is, mutual test environments deals with maximum number of sensors while compared to traditional IoT environments. So, the Traditional data mining techniques have to be

customized. These types of noises are not handled by the traditional data mining methods as these noises are not present in traditional IoT applications. Traditional IoT data analytics techniques are not suitable for mutual test environments. So, there is a need for new preprocessing techniques because the processing time is increased and Classifier accuracy is decreased by the presence of noise, sensor errors and irrelevant features. Table 3.2 shows the details of collected data.

Table 3.2: Dataset Description

Type			**Total**		
Total Data (Rows)			**3,19,520**		
Noise			**21,393**		
Point Noise	**Continuous Noise**		**13,856**	**7,537**	
Repetitive	**Collision**	**Null**	**16,596**	**2,653**	**1,548**
Sensor Errors			**596**		
Features (Columns)			**32**		

3.5. AIM AND OBJECTIVES

3.5.1. Aim

The proposed research work aims to propose an IoT data pre-processing architecture to clean the noisy data, to handle sensor errors and to select the best features to enhance the accuracy of IoT data analytics.

3.5.2. Objectives

- To propose DaRoN Technique to remove the noise in IoT data to improve the classifier accuracy.
- To propose TANOS Technique to remove the sensor errors in IoT data to attain a high accuracy rate.
- To propose MESIA Technique to select the best features in the collected dataset for pre-processing IoT data efficiently.

3.6. METHODOLOGY DIAGRAM

Jo's architecture consists of three phases, namely, the data collection phase, preprocessing phase and classifier phase.

3.6.1. Data Collection Phase

The data collection phase is the initial phase, which deals with data collection. Irrigation data were collected under a mutual test IoT environment by using five sensors: temperature sensor, soil moisture sensor, anemometer sensor (wind speed sensor), humidity sensor and rain sensor. Five sets of these sensors were placed in five different locations in the field. The mutual test environment has two types of sensors: parent and child sensors. The parent sensor is well placed and it has a low noise and error rate. Child sensors are supported sensors to the parent sensor for data collection. The parent sensor is placed in the center of the location and child sensors are placed in 4 corners of the field. This mutual test data collection process eliminates the blind spot problem in the data collection and makes the application suitable for real-time. Details of sensors are shown in table 3.3.

Table 3.3: Sensor Details

Sensor Name	Elements	Parent	Child	Unit
Temperature (T)	$T = \{t_1, t_2, t_3, t_4, t_5\}$	t_5	t_1, t_2, t_3, t_4	°C
Soil Moisture (S)	$S = \{s_1, s_2, s_3, s_4, s_5\}$	s_5	s_1, s_2, s_3, s_4	% Percentage
Rain (R)	$R = \{r_1, r_2, r_3, r_4 r_5\}$	r_5	r_1, r_2, r_3, r_4	Sensitivity Value
Humidity (H)	$H = \{h_1, h_2, h_3, h_4, h_5\}$	h_5	h_1, h_2, h_3, h_4	$g.kg^{-1}$
Anemometer (Wind Speed) (W)	$W = \{w_1, w_2, w_3, w_4, w_5\}$	w_5	$w_1, w_2, w_3, w_4,$	FPM
***FPM = Feet Per Minute**				

Sensor sets are defined as follows,

$T = \{t_1, t_2, t_3, t_4, t_5\}$ → **Temperature Sensor Values**

$S = \{s_1, s_2, s_3, s_4, s_5\}$ → **Soil Moisture Sensor Values**

$H = \{h_1, h_2, h_3, h_4, h_5\}$ → **Humidity Sensor Values**

$R = \{r_1, r_2, r_3, r_4\ r_5\}$ → **Rain Sensor Values**

$W = \{w_1, w_2, w_3, w_4, w_5\}$ → **Anemometer Sensor Values**

Therefore, L can be written as L = {T, S, H, R, W}

Location sets are defined as follows,

$L_1 = \{t_1, s_1, h_1, r_1, w_1\}$ → **Child Sensor Values**

$L_2 = \{t_2, s_2, h_2, r_2, w_2\}$ → **Child Sensor Values**

$L_3 = \{t_3, s_3, h_3, r_3, w_3\}$ → **Child Sensor Values**

$L_4 = \{t_4, s_4, h_4, r_4, w_4\}$ → **Child Sensor Values**

$L_5 = \{t_5, s_5, h_5, r_5, w_5\}$ → **Parent Sensor Values**

Therefore, L can be written as $L = \{L_1, L_2, L_3, L_4, L_5\}$

This type of data collection increases the data quality while comparing it with traditional data collection. Mutual test data collection partially eliminates missing values and outliers. But there exist few problems like sensor errors, repetitive values, null values, etc. Sensor sets and sensor set positions in various locations are shown in figure 3.4 and figure 3.5. In figure 3.4, parent sensor is placed in the center of the field and it has a wide sensitivity range but it has some blind spots. The child sensors are placed in four corners and it removes the blind spots of the parent sensor. So, both child and parent sensors are reliant on each other.

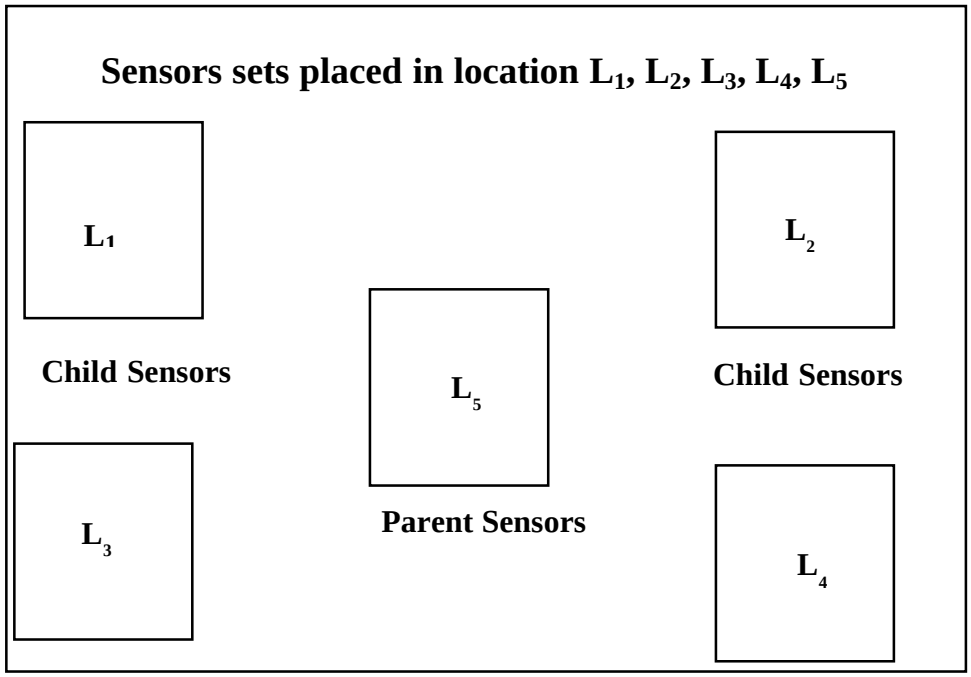


Figure 3.4: Data collection in the proposed Jo's architecture

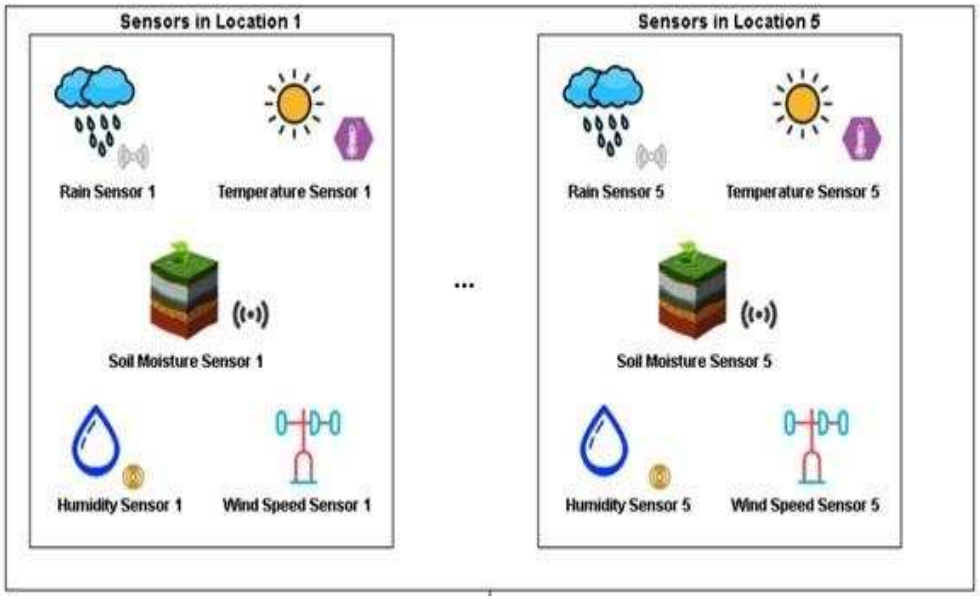


Figure 3.5: Sensor sets in various locations

Detailed descriptions of parent and child sensors are listed in table 3.4.

Table 3.4: Parent and Child Sensor details

Details	Parent	Child
Role	Master	Slave
Position	Well placed in the center	Placed in corners
Sensitivity range	Maximum	Minimum
Maximum No. of Sensors	Only One parent	Child = n
Work Done	Collect data without outlier and missing values	Collect data where the parent has a blind spot.
Selection	Most of the time parent sensor value is selected	Child sensors mean/ median/ mode values are selected when the parent has noise or sensor errors.
Data quality	High	Medium
Sensor errors	Low	High
Outliers	None	None
Data Size	Low dimension	High Dimension
Replacement value	Parent value or mean/ mode/ median values which are near to the parent are selected.	Parent value or mean/ median/ mode value of all sensors are selected in case of parent sensor has an error.

3.6.2. Preprocessing Phase

The proposed preprocessing phase consists of three techniques:

(i) A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency (DaRoN)

(ii) Technique for hAndling sensor errOrs in Smart irrigation system (TANOS)

(iii) enseMble filtEr-based feature Selection for IoT Agriculture Data (MESIA)

Detailed descriptions of DaRoN, TANOS, and MESIA are discussed in Chapters 4, 5, 6.

a. A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency (DaRoN)

In this proposed work, data is collected in the IoT irrigation environment under the mutual test category. So there exists noise in the form of repetitive values, point noise, continuous noise, class noise, attribute noise and collision values which are not handled by the existing techniques. Generally, the noise removal technique follows three stages of data processing: robust (detection of any analysis errors to make the data standardized), filtering (using various measures to remove noise) and polishing (Replacing error values). Each stage requires separate techniques. So, the proposed DaRoN technique, combines these three stages into a single step by using the timestamp value of sensors and central tendency measures. By using timestamp value, robust and filtering are done and after that polishing is done by using the central tendency measure. Best values are selected in a central tendency for polishing. Here, the best value is selected by the comparison with the reference sensor (Parent sensor) value.

In the mutual test environment, a sensor that is placed perfectly in a good position and has less possibility of noise, missing values and outliers are called reference sensors and others are called child sensors. DaRoN compares the central tendency value with the reference sensor value out of which the nearest value to the reference sensor value is selected for polishing. In the case of a sensor, errors are present in reference sensor value or child sensor value than that particular attribute is ignored. Sensor errors will be handled in the TANOS technique, after removing noise from the collected data.

b. Technique for hAndling sensor errOrs in Smart irrigation system (TANOS)

Outliers do not directly exist in the proposed environment, but missing values are present in the form of sensor errors. Sensor error occurs when the sensor fails to collect data. There are various reasons behind sensor errors such as, connection failure, power failure and sensor failure. The DaRoN technique removes the noise but ignores the sensor errors because error handling is easy when the dataset is balanced. If a data set has an equal number of elements in target classes, it is called a balanced dataset otherwise, it is called as imbalanced. After the DaRoN technique is applied in the collected dataset it is balanced. The missing value is categorized based on the position of attributes, which means all missing values are not harmful. The position of the missing values plays a vital role in categorization. Generally, missing values are classified into three types. They are Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). Among these, MAR and MCAR are not harmful but, MNAR is harmful which is handled in the proposed environment. Deletion is the best method for the removal of MAR and MCAR. To improve classifier accuracy, neighbor value and neighbor mean values are used to replace the missing values. Not a Number (NaN) error exists in the proposed environment which is removed by using neighbor value. If the error is found in the central sensor, the mean value of neighbor is used for replacement. By doing this replacement, the TANOS technique removes sensor errors. Sensor error is removed from the data set and is fed to the Support Vector Machine (SVM) classifier to check the accuracy. Finally, the TANOS technique is compared with existing error removal techniques and yields high rate of accuracy than others.

c. enseMble filtEr-based feature Selection for IoT Agriculture Data (MESIA)

MESIA Technique ensembles the univariate and multivariate filter methods by using mean and threshold values as supporting factors. Generally, the filter method requires additional support to identify the variable dependency. So, mean and threshold values are used in MESIA. Initially, Multivariate filtering is performed in MESIA for which mean value is calculated and subsets are selected. After selecting the best subset, environmental based threshold values are used to perform univariate filtering to eliminate the irrelevant features. Positive (ρ) and negative (-ρ) correlations are calculated in the univariate filter which eliminates irrelevant features based on environmental conditions. By using these 5 subset features, 5 different sensors are combined into one. This process enhances the classifier accuracy and reduces the machine learning model building time. The proposed technique is compared with existing techniques and is applied to the classifier to check the accuracy. The proposed technique proves that the accuracy of the classifier and the training time is reduced by using mean and threshold values.

3.7. WORKING OF Jo's ARCHITECTURE

Jo's architecture is proposed for the IoT data collected in a mutual test environment. The Architecture includes various phases such as the data collection phase, preprocessing phase and SVM classifier phase. The pre-processing phase combines the proposed three techniques such as Technique for Detection and Removal of Noise in IoT Data by using Central Tendency (**DaRoN**), Technique for hAndling sensor errOrs in Smart irrigation system (**TANOS**) and EnseMble FiltEr Based Feature Selection for IoT Agriculture Data (**MESIA**). Jo's architecture utilizes

these three techniques to provide Noise and Error free dataset for the analytical model. The **DaRoN** technique can be used for removing all types of noises in an IoT irrigation environment which comes under the mutual test category. If there are sensor errors (missing values), which is ignored by the DaRoN technique then the **TANOS** technique handles the sensor errors by using neighbor values replacement based on the error position. **MESIA** technique is used for selecting the appropriate features for building a machine learning model. Jo's architecture is shown in figure 3.6 and Figure 3.7 shows the real time data collection scenario.

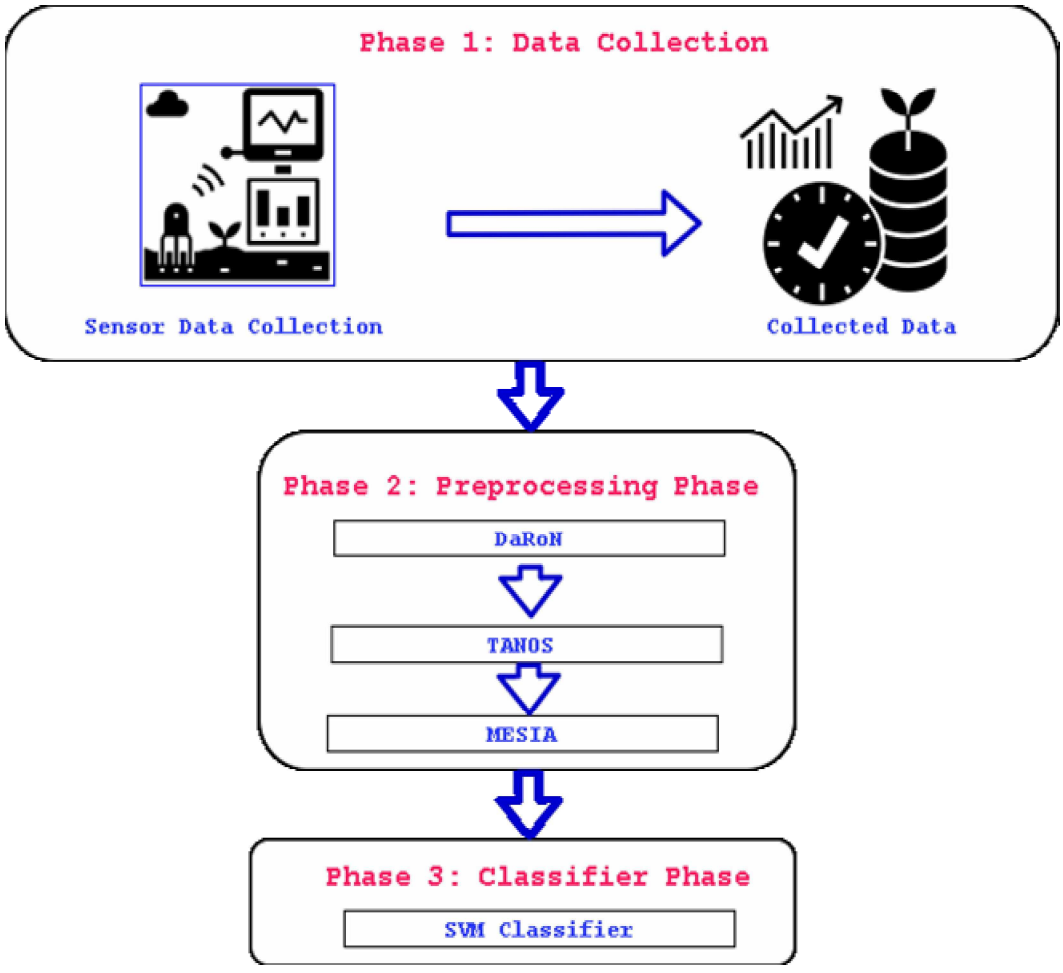


Figure 3.6: Proposed Jo's Architect

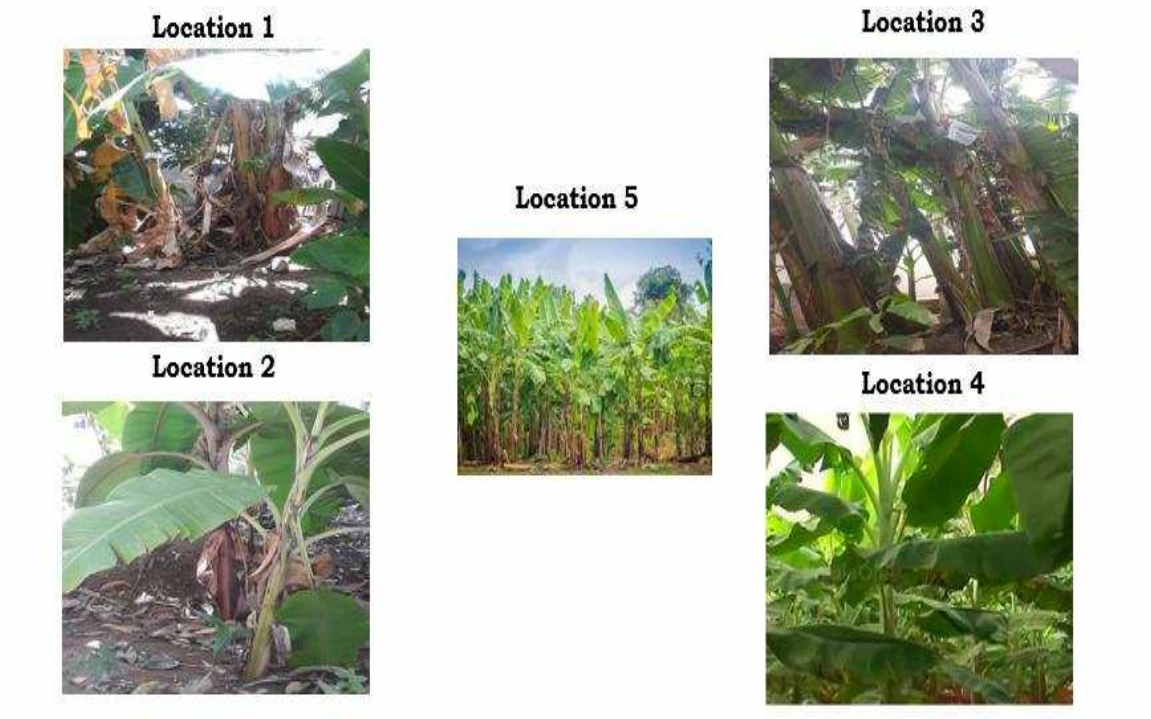


Figure 3.7: Data Collection Scenario

Table 3.5 shows the sensors used in the Jo's architecture. Figure 3.8 shows the sensor used for data collection. Table 3.6 shows the system requirements used for Jo's architecture.

Table 3.5: Sensors used in Jo's architecture

Sensor Name	Model Number
Humidity sensor	DHT11
Soil Moisture Sensor	RC-A-4079
Rain Sensor	Y2-LLZY-30GH
Temperature sensor	LM35
Anemometer	WS102
Bluetooth Module	HC-05

Table 3.6: Configuration of the system

Type	**Specification**
Processor	i7-7500U
CPU	2.70GHz
Random Access Memory (RAM)	12GB
Programming Language	Python with NumPy, Pandas, and Tensor Flow Libraries
Arduino Uno	ATmega328P – 8 bit AVR family microcontroller (5 Nos.)
Humidity Sensor	DHT11 (5Nos.)
Soil Moisture Sensor	RC-A-4079 (5Nos.)
Rain Sensor	Y2-LLZY-30GH (5Nos.)
Temperature sensor	LM35(5Nos.)
Anemometer	WS102 (5Nos.)
Bluetooth Module	HC-05 (5Nos.)

Figure 3.8: Sensors used for Data collection

3.7.1. Procedure for Jo's Architecture

(i) Procedure of DaRoN Technique

Step 1: Time stamp value is used to remove the continuous noise and repetitive values.

Step 2: Point noise and Error value removal using central tendency.

Step 3: Null value Removal using central tendency.

(ii) Procedure of TANOS Technique

Step 1: Assigning the neighbor for each sensor.

Step 2: Child sensor error removal by using neighbor value or parent value.

Step 3: Parent sensor error Removal by using the mean value of child sensors.

(iii) Procedure of MESIA Technique

Step 1: Multivariate Filtering is done by using central tendency.

Step 2: Univariate filtering is done by using seasonal based threshold values

Step 2.1: Condition for the rainy season.

Step 2.2: Condition for Cold Season.

Step 2.3: Condition for Hot Season

Step 2.4: Condition for Strom Season

Step 2.5: Condition for Cloud mask Season

3.8. CHAPTER SUMMARY

This chapter discussed the proposed architecture for IoT-based smart irrigation system. The proposed architecture aims to clean the collected IoT sensor data. As the collected dataset has many defects such as noise, sensor errors, outliers etc., it is not suitable for decision making. The proposed Jo's architecture contained three prominent

phases, namely data collection phase, preprocessing phase and classification phase. In which the mutual test environment is developed for data collection. The data is collected from various agricultural related sensors from different locations for irrigation. 25 sensors are placed in five different locations for this purpose. After data collection, the preprocessing phase is enabled. In this phase, three different techniques are proposed to clean the data. They are **D**etection **a**nd **R**emoval **o**f **N**oise in IoT Data by using Central Tendency **(DaRoN), T**echnique for h**A**ndling se**N**sor err**O**rs in Smart irrigation system (**TANOS**) and ense**M**ble filt**E**r-based feature **S**election for **I**oT **A**griculture Data **(MESIA)**. DaRoN is proposed to detect and remove the noise in the collected sensor data, TANOS is proposed for handling the sensor errors in the collected set and MESIA technique is proposed for selecting the relevant features for decision making during various seasons like rainy season, cold season, summer, etc. Eventually, all these three techniques are combined under Jo's Architecture which selects the appropriate technique according to the data cleaning problem. The third phase of Jo's architecture is classification. The preprocessed data is given to this phase and SVM classifier is utilized. Based on the classification results, the decision is made. All the proposed three preprocessing techniques are evaluated in terms of accuracy, precision, recall and F1 score using the confusion matrix and result in a high accuracy rate. Hence, Jo's architecture handled the mutual test IoT data perfectly and improved the classification accuracy.

Chapter – 4

DaRoN: A Technique for Detection and Removal of Noise in IoT Data By Using Central Tendency

Chapter – 4

DaRoN: A Technique for Detection and Removal of Noise in IoT Data By Using Central Tendency

4.1. INTRODUCTION

IoT is the supreme technology for data collection. In this DaRoN technique, data are collected using IoT sensors under mutual and multi-test environments. A mutual test environment has the same set of sensors used in the same field for data collection whereas a multi-test environment has a different set of sensors [**Zho, 19**]. Traditional data collection problems are avoided by using a mutual and a multi-test environment, but there arises new problems like noise and sensor errors. Sensor errors are caused by device failures, interrupted connection etc. For example, Not a Number (NaN) error is a type of sensor error caused by interrupted connection. Noisy data are meaningless or useless. Noisy data need to be preprocessed before processing [**Ass, 17**]. Because, noisy data cause two major problems in data analytics. One is an increase in model training time and the other is unreliability in classification and prediction accuracy. Noise in data makes the process of extracting meaningful information a tedious one. So, a novel technique DaRoN is proposed for noise detection and removal of the data collected from IoT sensors under mutual test and multi test environments.

4.1.1. Categories of Noise

In the collected data, many forms of noise can be present like repetitive values, null values, error values, missing values, outliers, sensor errors etc. Noisy data can be categorized based on two aspects **[Mor, 19]**. One is based on the position of

the noisy data and another is based on the type of noise present in the data. Position based noisy data can be further classified as class noise and attribute noise. Type based noisy data can be classified as point and continuous noise. The categorization of the noisy data is depicted in figure 4.1.

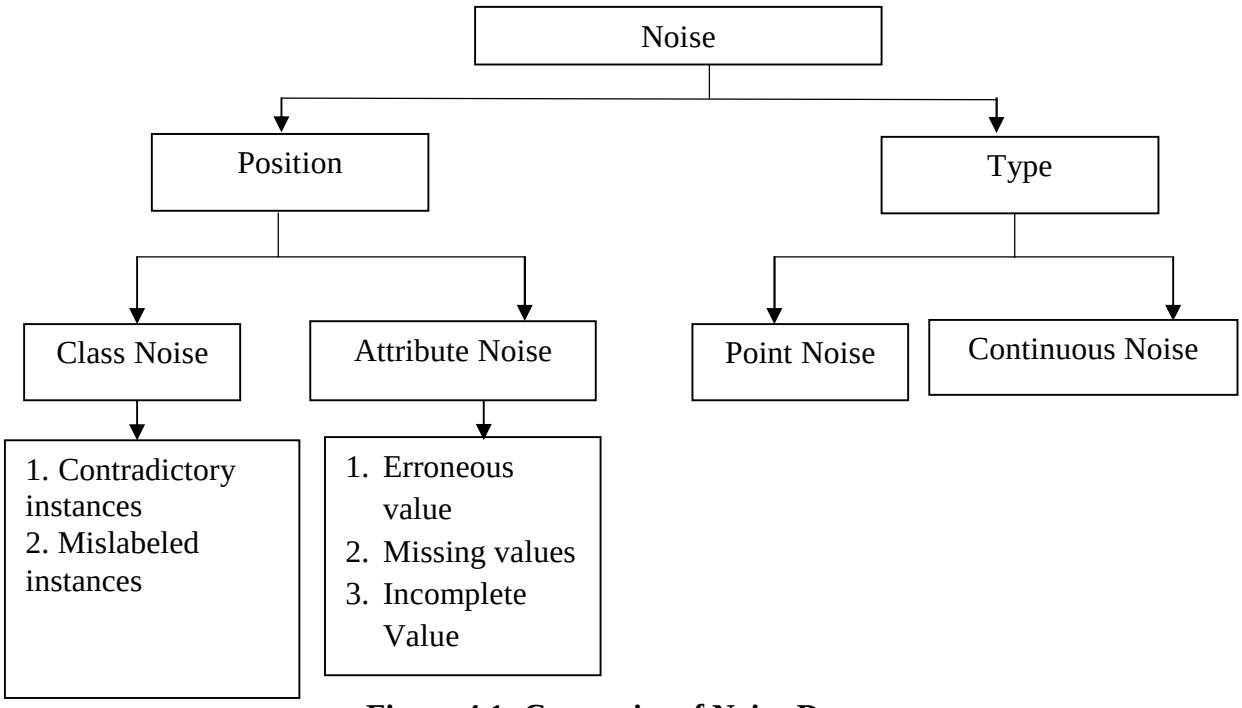


Figure 4.1: Categories of Noisy Data

a. Class Noise

Class noise is a type of noise that occurs when noise is present in the class label. Class noise removal is a tedious task for all data analysts because these noises cause ambiguity whether all instances of a class are properly classified under the appropriate class label or not [**Zho, 19**]. There are two types of class noise such as Contradictory instances and Mislabeled instances. Contradictory instances are said to occur when the same attribute value appears in different class labels. *For* example, temperature sensor 1 and humidity Sensor 1 are defined as t_1 and h_1 respectively. The class label is positive for attributes such as temperature (t_1) = 35, humidity (h_1) = 40

and the class label is negative for one instance and positive for another instance. Then it is called as contradictory instance.

The mislabeled instance is said to occur when attribute values are misplaced under the wrong class label. This error is common in data analytics. For example, the class label is positive for attributes such as temperature (t_1) = 18, humidity (h_1) = 32 and the class label is positive for t_1= 38 and h_1 = 59 that is the same attributes with different values.

b. Attribute Noise

Attribute noise occurs when noise exists in the attribute value. Attribute noise is classified into three types namely erroneous value, missing value and incomplete value **[Gar, 18]**. Comparing with class noise, attribute noise is more harmful because, as it directly affects the prediction results. Erroneous values are stored rather than the measured sensor value or actual value. For example, the actual measured temperature sensor value (t_1) 40º C may be wrongly stored as 20 º C. If the decision is made based on the stored wrongly data, then the classifier accuracy will not be reliable. Missing Values occur when data collection is interrupted due to poor network connection or power failure or environmental issues like earthquakes, storms etc., In these cases, there is a possibility of data loss and without preprocessing these data, one cannot proceed further with the analysis.

For example, consider a smart agricultural environment. If the humidity sensor data (h_1) is only available for 2 hours for a day then no decision can be taken and it may lead to ambiguity. Generally, it is better to handle missing values after cleaning class noise and attribute noise. Incomplete value means completeness of data is not

available. For example, in smart irrigation, decisions will be taken based on the combination of two or three sensor data (soil moisture (s_1) and humidity sensor (h1) data). In this case, if anyone sensor data is not available then the decision making process is affected. Because of this reason, attribute noise is considered more harmful than others. Types of noise are shown in table 4.1 with examples.

Table 4.1: Types of Noise

Temperature sensor Value (t_1) (^oC)	**Humidity sensor value (h_1) ($g.kg^{-1}$)**	**Target Class**	**Types of Noise**
35	40	Wet (Positive)	Contradictory instances
35	40	Dry (Negative)	
18	32	Wet (Positive)	Mislabled instances
38	59	Wet (Positive)	
15	23	Wet (Positive)	Erroneous value
95	145	Dry (Negative)	
	23	Dry (Negative)	Missing values
18	0.00	Wet (Positive)	Incomplete Value
**Noisy data are highlighted in table*			

c. Point Noise

The Point noise arises when there are sudden deviations in data points, this can be easily identified **[Pet, 17]**.

d. Continuous Noise

Continuous noise occurs when there is a gradual rise in the data points and the deviation is difficult to be identified [**Jam, 19**] [**Asi, 18**]. Pictorial representation of point noise and continuous noise is depicted in figure 4.2.

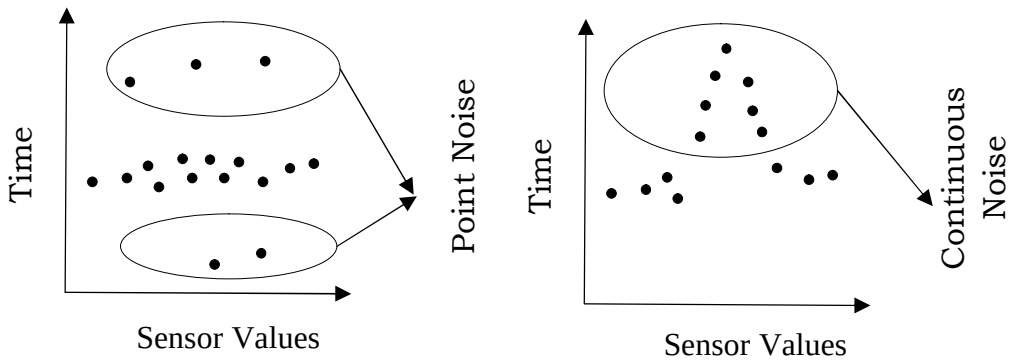


Figure: 4.2: Types of Noise Data

4.2. BACKGROUND STUDY

4.2.1. Noise removal techniques

Noise removal techniques [**Liu, 20a**] are classified into three types. They are: feature selection methods, techniques to remove noise in records and anomalies based methods. Feature selection methods [**San, 18**] are used to select and eliminate features from the collected data. Generally feature selection methods are best for removing noise in columns. Feature selection methods are classified into three types which are filter method, wrapper method and embedded method. The popular method used for noise removal in records is the K fold validation [**Cho, 01**] and manual validation. K fold validation is not suitable for streaming data (IoT data) and large data. Manual methods are used for special datasets such as mutual test data (Collected data environment). This method adapts various techniques to remove noise which is suitable for the collected data set. Anomaly detection is used to identify the deviated data from the collected data [**Lan, 17**]. So this method is suitable for outlier detection and removal. Here, Data is collected in a mutual test environment. So, traditional noise removal method does not apply to other environments. Noise removal techniques are depicted in figure 4.3.

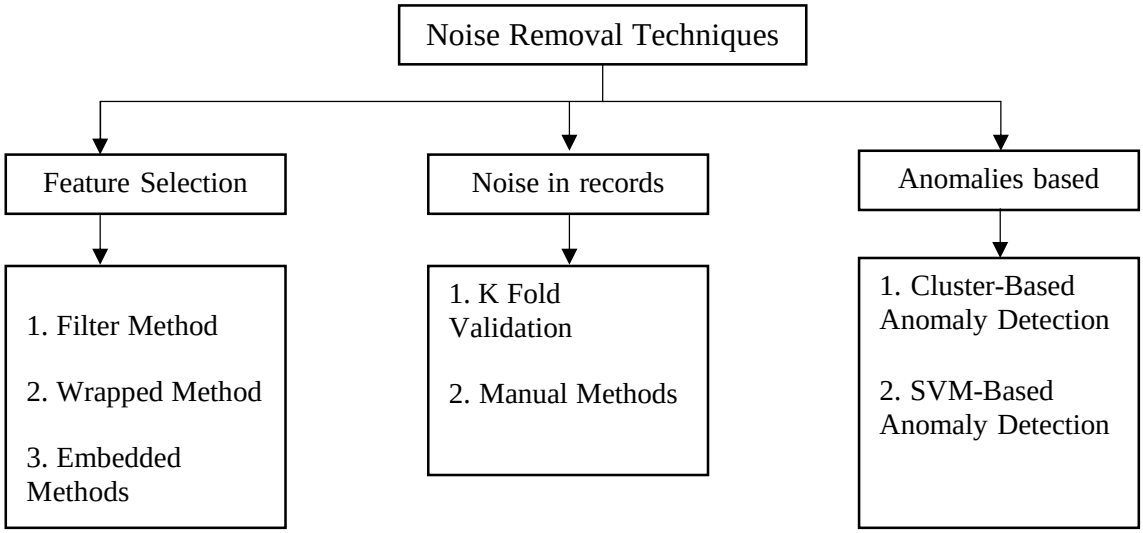


Figure: 4.3: Traditional Noise Removal Techniques

4.2.2. Working with noise removal techniques

All noise removal techniques undergo three stages. They are: robust, filtering and polishing. Robust is used to remove errors and make the data standardized. Filtering is used to remove noise or to select noiseless data. Polishing is used to replace the error values. These three stages require separate techniques for the processing which leads to an increase in time and reduces the accuracy of the model.

4.3. RELATED WORKS

Garcia et al., [**Gar, 16**] proposed a method to improve the accuracy of the noise removal method. Meta Learner (MTL) was used for removing redundant data and irrelevant data. Meta features were used for creating new features from the corrupted features. Class noise was not handled properly in this method so system processing time was increased.

Saez et al., [**Sae, 17**] proposed a technique to detect and filter noisy data. Iterative Class Noise Filter (INFFC) was used for noise filtering, which is done in three stages namely Preliminary Filtering, Noise-free filtering, and Final removal of

noise. Ensemble Filter (EF) used multi classifiers: Support Vector Machine (SVM), C4.5, K Nearest Neighbor (KNN)) to remove noisy data. The noise was removed iteratively by using Iterative-Partitioning Filter (IPF). This technique did not handle attribute noise and continuous noise.

Wang et al., [**Wan, 20**] proposed a framework to increase prediction accuracy by removing noise in the wind data. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEDMAN) technique was used for noise removal. Multi-tracker Optimizer (MTO) was used for noise detection. This method was only suitable for small data set because mean error increased while using a large dataset. Existing noise removal methods are listed in table 4.2.

Table 4.2: Existing Noise removal Methods

Author Name	Robust	Filter	Polish
[**Gar, 16**]	MTL	MTL	Meta features
[**Sae, 17**]	INFFC, EF, IPF	INFFC, EF, IPF	Not considered
[**Wan, 20**]	CEEDMAN	MTO	Neural network

4.4. Need for research

Repetitive values, null values, error values and sensor errors affect the collected data set. The literature indicated repetitive values, null values, error values and sensor errors have not been handled properly by the existing techniques**.** Figure 4.4 shows the details of collected data under a mutual test environment.

There is a significant difference between the traditional IoT environment and the Mutual test environment. Problems like repetitive values and error values are highly present in the mutual test environment. So traditional preprocessing methods are not suitable for this data. Figure 4.5 shows that the performance of traditional noise removal techniques testing with the collected data.

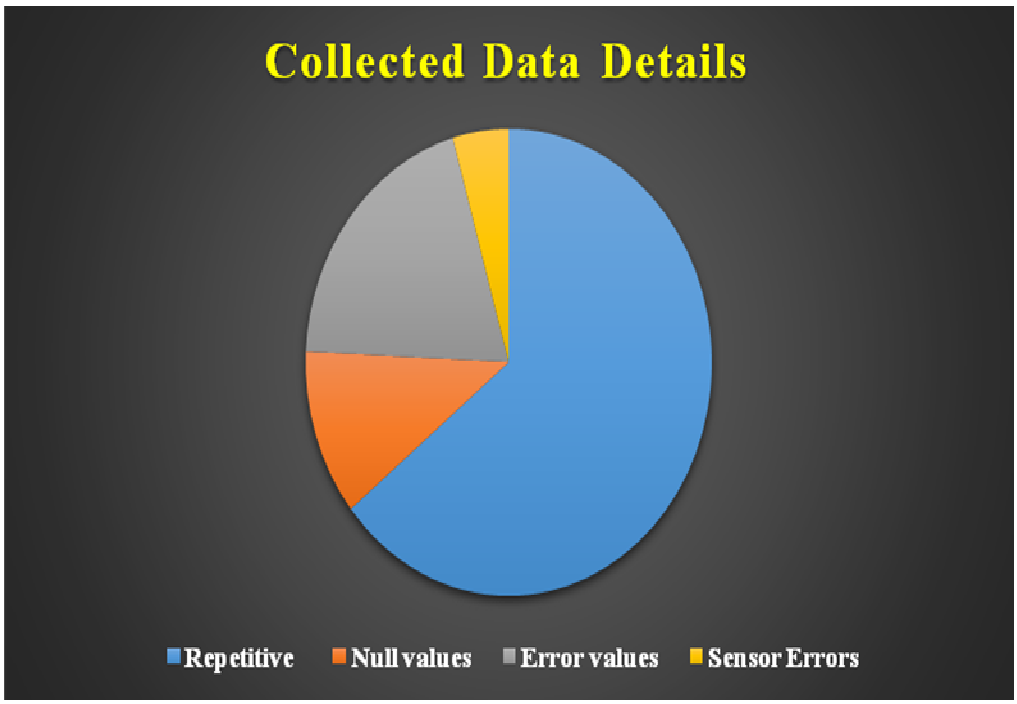


Figure 4.4: Type of Noise in collected data

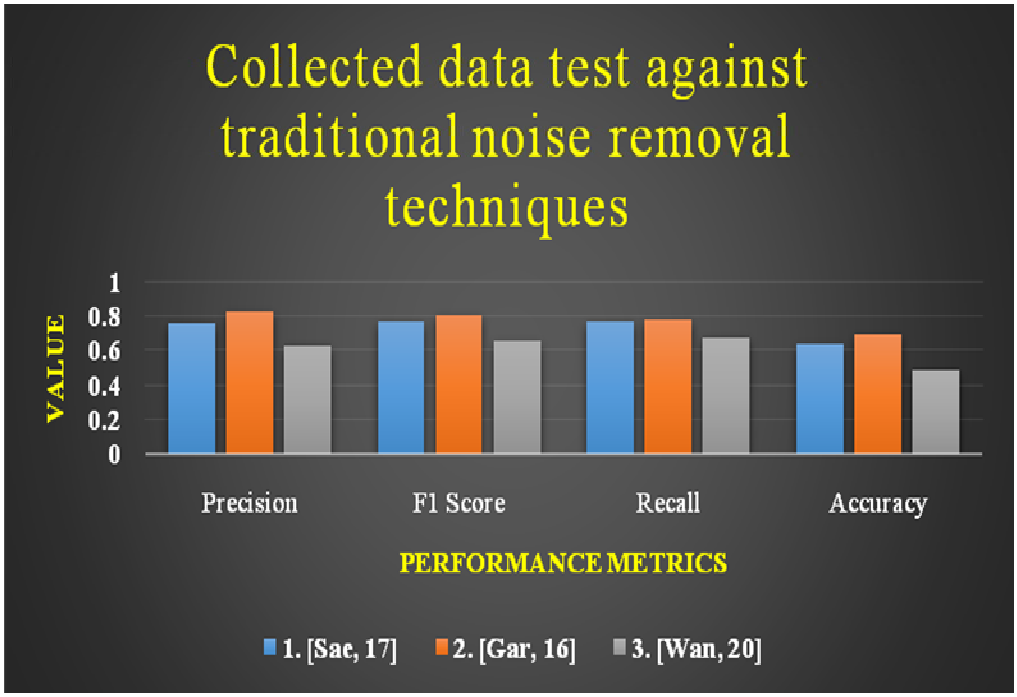


Figure 4.5: Traditional noise removal technique performance

Figure 4.5 proves that traditional noise removal methods are not capable of handling the mutual test environment. The mutual test environment is a growing research area in the field of data preprocessing. Hence, this research work focuses on the Noise removal in mutual test environment.

4.5. OBJECTIVES

The chapter aims to detect and remove noise in order to improve the accuracy of decision making. Repetitive values, null values, error values and large data sets agonize the existing methods a lot. This has motivated to propose a new technique to handle large data set, repetitive values, null values, error values and all types of noise.

- To handle types of noise (point noise, continuous noise, attribute noise and class noise) separately.
- To eliminate the repetitive, null and error values without corrupting the collected data.
- To combine the traditional stages of noise removal process.

4.6. METHODOLOGY DIAGRAM

The proposed DaRoN technique combines various stages of traditional noise removal techniques that is robust, filtering and polishing. IoT sensor has timestamp value ie., Time details of the data collected which is used for robust and filtering. Mean values replacement is used for polishing. Existing methods require a separate technique for robust, filtering and polishing which lead to an increase in processing time and system complexity.

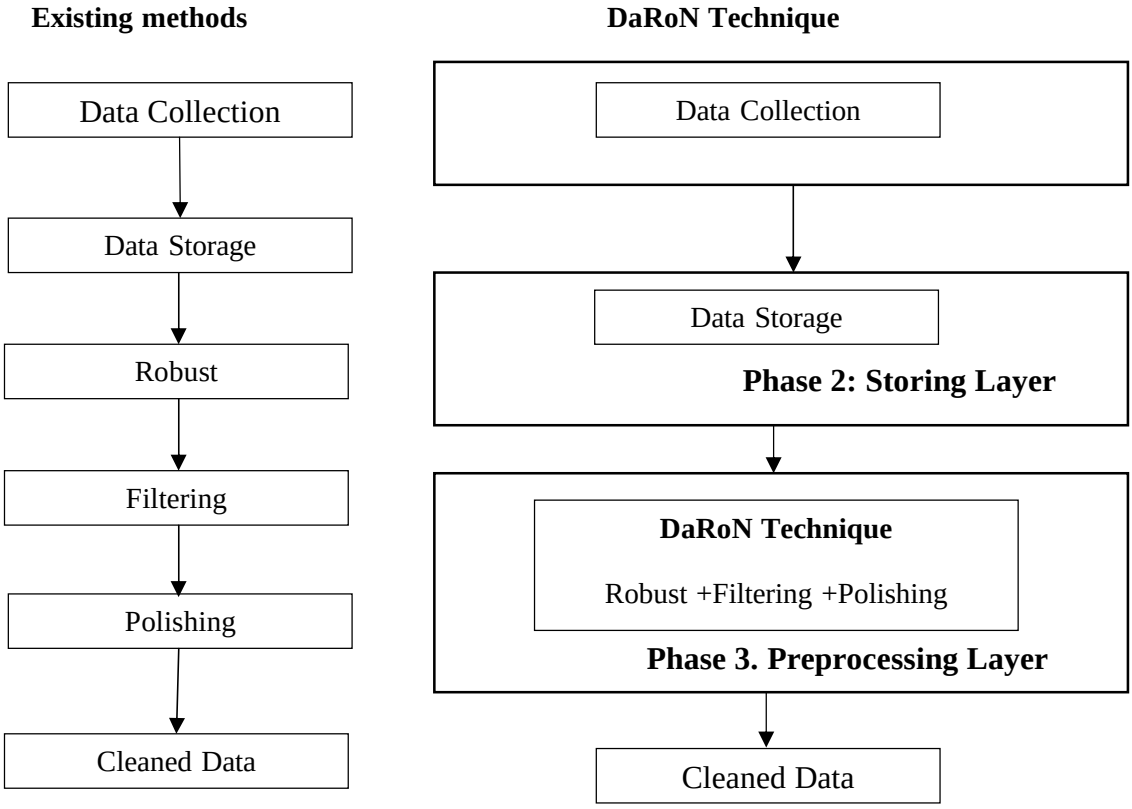


Figure 4.6: Methodology of the Existing vs. DaRoN technique

All types of noise are not handled properly by the existing methods. So accuracy is not reliable. Thus, it is not suitable for real time but, DaRoN handles all types of noise by combining the three stages of noise removal. DaRoN methodology is shown in figure 4.6.

The data is collected under a mutual test environment. So, traditional mean value is used for polishing. In a mutual test environment, general noises like missing values and outliers are avoided. But, there is a possibility of other forms of noise. They are repetitive values, error values, null values, sensor errors etc. Sensor error occurs when a sensor failed to collect data generally sensor errors are existed in the form of missing values. Sensor error will be handled after balancing the data. There are two types of sensors in the mutual test environment. One is the parent sensor

which is perfectly placed and has less noise. So, noises are easily removed by using parent sensor values. The second type of sensor is the child sensor which is placed near to parent and plays a supporting role to the parent sensor. The mutual test environment is shown in figure 4.7.

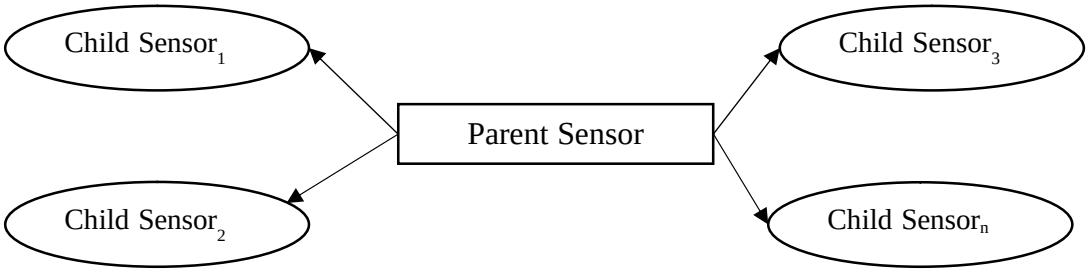


Figure 4.7: Mutual test environment

4.6.1. Working of DaRoN technique [Jan, 21b]

Data are collected by using 5 sets of sensors (Soil Moisture sensor (S), Temperature Sensor (T), Humidity Sensor (H), Rain Sensor(R), and Wind speed sensor (W)) which are placed in 5 different locations. So totally there are 25 sensors. T, S, H, W and R represent a sensor set each has 5 elements (members) listed below.

$T = \{t_1, t_2, t_3, t_4, t_5\}$,

$S = \{s_1, s_2, s_3, s_4, s_5\}$,

$H = \{h_1, h_2, h_3, h_4, h_5\}$,

$R = \{r_1, r_2, r_3, r_4 r_5\}$

$W = \{w_1, w_2, w_3, w_4, w_5\}$

$L_1 = \{ t_1, s_1, h_1, r_1, w_1\}$ similar for L_2, L_3, L_4, L_5

Therefore, L can be written as L = {T, S, H, R, W}

Initially, robust and filtering is performed based on the timestamp value. That is, one observation for every two hours. This reduces the 1440 data readings (24 hours

data) into 12 data readings per day. The continuous noise is partially avoided by using robust and filtering because generally continuous noise is identified after reaching a certain point. But, the proposed condition splits the data into 12 observations per day so that continuous noise is partially avoided. After polishing, continuous noise is entirely handled. Data with the same timestamp are removed so that repetitive values are filtered. Data are selected on a regular intervals of 10 minutes. For example, from the filtered data, 10^{th}-minute data is selected for day 1; 20^{th}-minute data is selected for day 2. This is done because, data collected at regular intervals will avoid continuous noise. Alternative minute selection is used to avoid continuous noise. This process is continued for the entire collected data. The approach used for data selection is shown in table 4.3.

Table 4.3: Data selection approach

Day 1	**Time Period**	**Anti Meridiem (AM)**						**Post Meridiem (PM)**					
	Hour	1-2	2-4	4-6	6-8	8-10	10-12	12-2	2-4	4-6	6-8	8-10	10-12
	Minutes	10	10	10	10	10	10	10	10	10	10	10	10
Day 2	**Time Period**	**Anti Meridiem (AM)**						**Post Meridiem (PM)**					
	Hour	1-2	2-4	4-6	6-8	8-10	10-12	12-2	2-4	4-6	6-8	8-10	10-12
	Minutes	20	20	20	20	20	20	20	20	20	20	20	20

Point noise which is in the form of error values is removed by comparing child sensor mean or mode or median values against parent sensor value. If the values are

near to the parent sensor value then it remains unchanged else it is replaced with the parent sensor value.

$$Mean\ (\mu) = \frac{sum\ of\ \ all\ elements}{Total\ number\ of\ \ Elements}$$

$$Mode\ (Z) = L + h\frac{((f_m - f_1))}{((f_m - f_1) - (f_m - f_2))}$$

$$Median\ (M) = \frac{(n+1)}{2}$$

If a sensor error occurs in any elements of R, T, S, W, H then it will be ignored. Because, the sensor error will be handled only after balancing the data set. Most of the time mode value is not near to the parent sensor value hence it is not selected. The mean, median and mode are calculated exclusive of r_5 because r_5 is the parent. (Common for all t_5, s_5, h_5, w_5). Mean, Median and Mode election are shown in table 4.4.

Table 4.4: Mean, Median, Mode selection example

r_1	r_2	r_3	r_4	r_{5**}	**Mean**	**Median**	**Mode**	**Selected**
318	429	589	651	520	496.75	509	589, 651, 318, 429	**Median**
257	284	304	369	314	303.5	294	284, 369, 257, 304	**Mean**
187	201	252	198	NaN*	-	-	-	**Ignored**
NaN	196	423	385	200	-	-	-	**Ignored**

*** Not a Number (NaN) sensor Error**
**** Mean, Median, Mode are calculated exclusive of r_5**
Nearest value to the parent is selected

Null values are replaced with the mean value of the rest of the elements. For example, if t_1 has a null value, it is replaced by the mean value of t_2, t_3, t_4, t_5. If no null values are found, this process is continued for all the elements in T, S, W, H, R. Table 4.5 shows the null value replacement procedure. Thus, DaRoN handled all noise.

Table 4.5: Null value replacement procedure

t_1	t_2	t_3	t_4	t_5	Mean Calculation	Mean Value*
25	26	0	29	24	$\frac{t_1+t_2+t_4+t_5}{4}$	26
28	29	28	31	0	$\frac{t_1+t_2+t_3+t_4}{4}$	29
0	0	34	36	30	$\frac{t_3+t_4+t_5}{3}$	33
T (t_1,t_2, t_3, t_4, t_5)values are °C scale * null values replace by mean values						

The steps involved in the DaRoN technique is discussed below,

4.6.2. Steps for DaRoN technique

Step 1: Data is collected using various sensors for every two hours from various locations

Step 2: Collected data is stored in cloud database

Step 3: Data is retrieved from the cloud for preprocessing

Step 4: Mean (), Median () and Mode () are calculated for all sensor data

Step 5: Each sensor value is compared with the next value using time value (Td)

Step 6: The redundant values are identified and removed

Step 7: After that, Point noise and error values are identified

Step 8: The identified point noise and error values are replaced by the computed Mean (), Median () and Mode ()

Step 9: Redundant values, point noise and error values are successfully removed

Step 10: Cleaned data is generated

4.6.3. DaRoN technique

Technique: DaRoN

Input: Raw data

Output: Cleaned Data

$L = \{L_1, L_2, L_3, L_4, L_5\}$ (or) $L = \{T, H, W, S, R\}$ // ***L denotes Location***

$L_1 = \{t_1, h_1, w_1, r_1, s_1\}$

$T = \{t_1, t_2, t_3, t_4, t_5\}$ // ***T denotes temperature sensor values***

$R = \{r_1, r_2, r_3, r_4, r_5\}$ // ***R denotes Rain sensor values***

$W = \{w_1, w_2, w_3, w_4, w_5\}$ // ***W denotes Wind sensor values***

$S = \{s_1, s_2, s_3, s_4, s_5\}$ // ***S denotes Soil Moisture sensor values***

$H = \{h_1, h_2, h_3, h_4, h_5\}$ // ***H denotes Humidity sensor values***

Initialize i = 0;
Set timer=00:00:00:00
while (i< 12) // ***One observation in every two hours***
    select R (Td[i]) // ***where R denotes Rain sensor values, R = {r_1, r_2, r_3, r_4, r_5}***
    select T (Td[i]) // ***where T denotes Temperature sensor values, T = {t_1, t_2, t_3, t_4, t_5}***
    select W (Td[i]) // ***where W denotes Wind sensor values, W = {w_1, w_2, w_3, w_4, w_5}***
    select H (Td[i]) // ***where H denotes Humidity sensor values, H = {h_1, h_2, h_3, h_4, h_5}***
    select S (Td[i]) // ***where S denotes Soil sensor values, S= {s_1, s_2, s_3, s_4, s_5}***
    sleep(7200) // ***7200 seconds***
  i++

if(i==12)

reset timer

i=0

compute R (μ), R (M), & R (Z); ***// mean, median and mode value of rain sensor***

compute T (μ), T (M), & T (Z); ***// mean, median and mode value of Temperature sensor***

compute W (μ), W (M), & W (Z); ***// mean, median and mode value of Wind sensor***

compute S (μ), S (M), & S (Z); ***// mean, median and mode value of Soil Moisture sensor***

compute H (μ), H (M), & H (Z); ***// mean, median and mode value of Humidity sensor***

for (int i=0; i<12; i++)

if(r_1 (Td[i]) < r_1(Td[i+1])) ***// Checking Redundant values based on time (Td)***

remove r_1(Td[i])

compute rest of R, and all elements in T,W, H, S

X = ((R≅R (μ), R (M), & R (Z))) ***// check approximate nearest value***

else if (X = = True)

compare all elements of R with R (μ), R (M), & R (Z) ***// select the nearest value***

replace with R (μ), R (M), & R (Z)

compute all elements in T,W, H, S

else if (r_1> 0)

keep the values

compute rest of R, and all elements in T,W, H, S

else

replace with R (μ), R (M), & R (Z)

end

end if

end for

DaRoN technique successfully handled all types of noise. Figure 4.8 shows the types of noise handled by DaRoN.

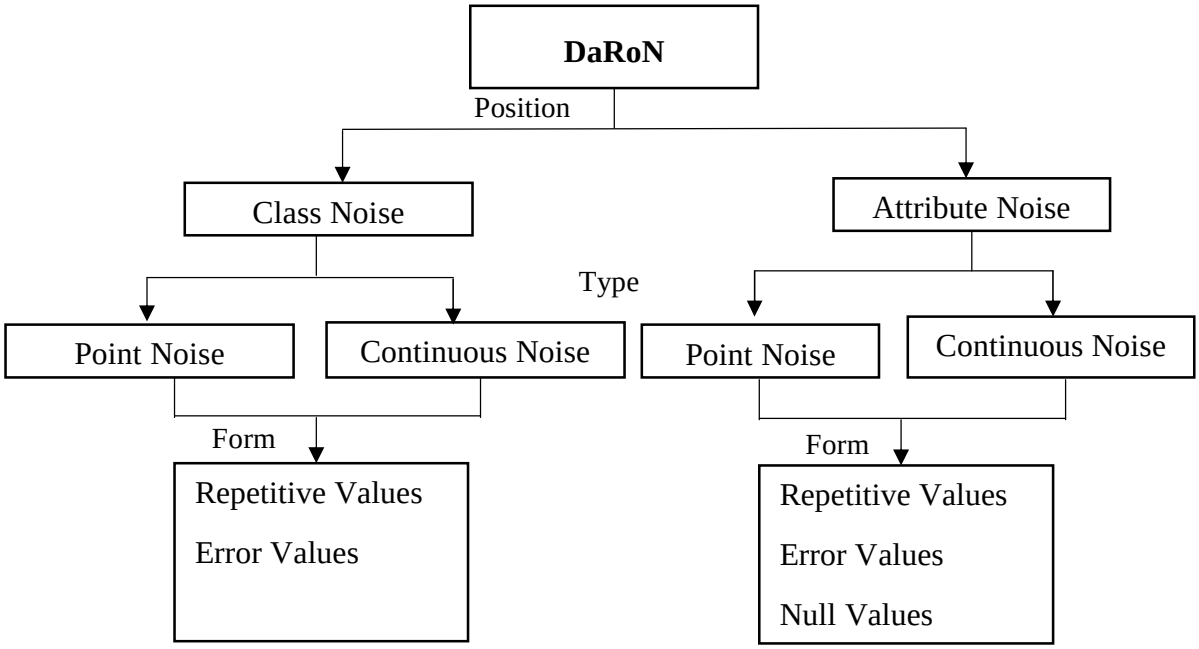


Figure 4.8: Noise handled by DaRoN

As depicted in figure 4.8, the proposed DaRoN Technique removed the noise in the sensor data.

4.7. RESULTS AND DISCUSSIONS

For this research work, data are collected in a mutual test environment. So, it cannot be compared to traditional methods. From the figure 4.5 it is clear that the existing techniques are not suitable for handling these data. The DaRoN technique is applied to the SVM classifier to check the performance. Confusion matrix metrics (precision, recall, accuracy, f1 score) are used to measure the performance of DaRoN. In figure 4.9 performance metrics chart, X-axis represents the performance metrics and the Y-axis represents the confusion matrix values. DaRoN achieved 96.05% precision i.e., positively predicted values from the collected data. F1 score is calculated

using precision and recall measures. DaRoN achieved 90.27% F1 score. The recall is the detection possibility of the classifier. DaRoN achieved 85.14 recall. Accuracy is the fraction between correct prediction and total prediction. DaRoN achieved 84.18% accuracy. Figure 4.9 shows the overall performance of the DaRoN technique.

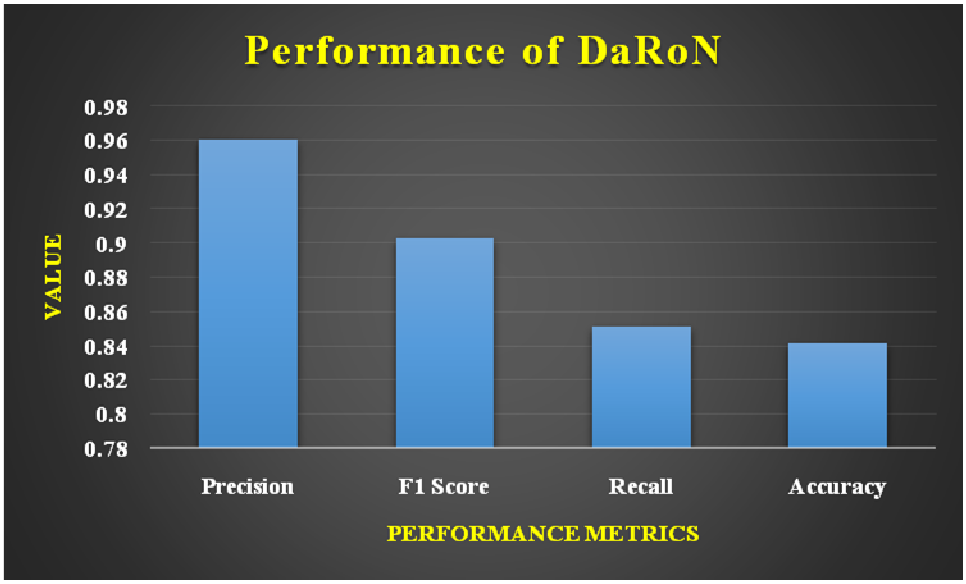


Figure 4.9: Performance of the DaRoN

4.8. FINDINGS AND INTERPRETATIONS

This section explains the experimental results of the DaRoN technique It explicates the strength of the DaRoN technique and how DaRoN achieves better results compared with traditional noise handling techniques. It also justifies how the DaRoN technique is better than the traditional noise handling techniques.

The DaRoN technique was used as the central tendency measure to remove noise in the mutual test IoT environment data. But DaRoN has not focused on sensor errors.

In this research, the DaRoN technique is applied to the SVM classifier to evaluate the performance of the technique. DaRoN achieved better performance than existing noise handling techniques in all performance metrics. So DaRoN improved

the SVM classifier performance better than existing techniques. Combining the three noises, handling stages into a single step by using timestamp value and central tendency method is the reason for the good performance of DaRoN technique. Table 4.6 shows the performance of DaRoN technique.

Table 4.6: Performance of DaRoN technique

Technique Name	Performance Metrics			
	Precision	**F1 Score**	**Recall**	**Accuracy**
DaRoN	0.9605	0.9027	0.8514	0.8418

4.9. CHAPTER SUMMARY

Traditional noise removal techniques have not considered all types of noise (Class noise, point noise, attribute noise and continuous noise). These methods are not capable of handling data collected under a mutual test environment. There are three stages in noise handling technique, namely robust, filtering and polishing. The proposed DaRoN used timestamp value for robust and filtering and central tendency measures for polishing. Combining the noise processing stages in the proposed DaRoN technique achieves better performance than the traditional techniques. DaRoN provided good results in terms of all performance metrics. DaRoN satisfied the objectives namely, to handle types of noise (point noise, continuous noise, attribute noise and class noise) separately, to eliminate the repetitive, null and error values without corrupting the collected data and to combine the traditional stages of the noise removal process. DaRoN handled all types of noise except sensor errors. After removing class noise, the data set is balanced. Only when the noisy data is removed sensor errors can be handled. So, sensor error will be handled in the next work.

Chapter – 5

TANOS: Technique for hAndling seNsor errOrs in Smart irrigation system

Chapter – 5

TANOS: Technique for hAndling seNsor errOrs in Smart irrigation system

5.1. INTRODUCTION

According to Yi et al., [**Yi, 19**], mutual test environments are not directly affected by missing values and outliers. Missing values are present in the form of sensor errors. For example, Not a Number (NaN), error Infinity (Inf) etc. NaN error exists in the collected data and is listed in table 5.1. These sensor errors occur due to connection failure, sensor failure and power failure. There are various reasons for sensor errors, which are listed in figure 5.1.

Table 5.1: Collected Data with sensor errors

Soil Moisture	Temperature	Humidity	Wind	Rain	Status
0.05	27	40	**NaN**	100	Low
0.06	27	**NaN**	210	100	Low
0.08	**NaN**	45	500	110	Medium

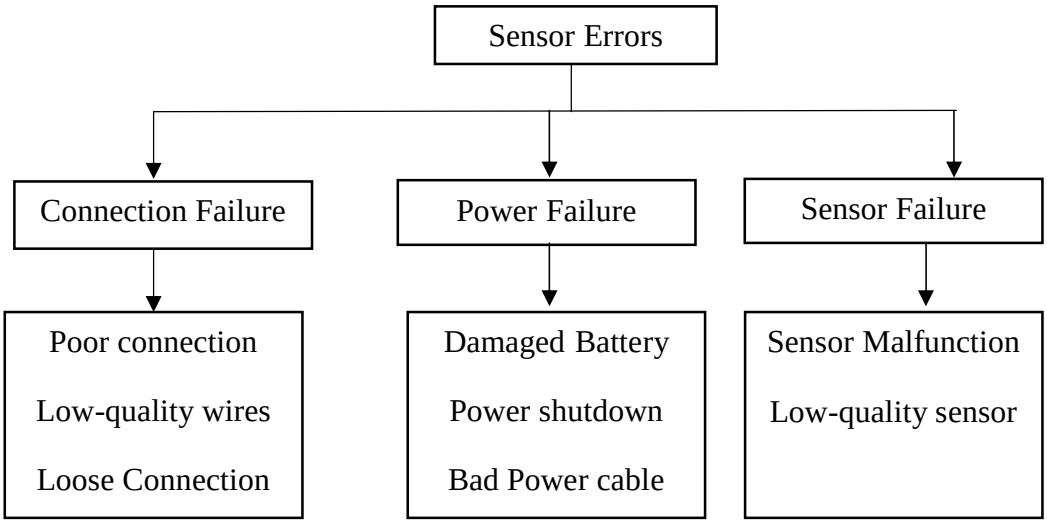


Figure 5.1: Reasons for Sensor Error

Generally, while collecting data, some rows are empty or valueless. These values are known as missing values. Missing values [**Yen, 19**] frequently exist in many IoT environments. The problems endured are mentioned in figure 5.2.

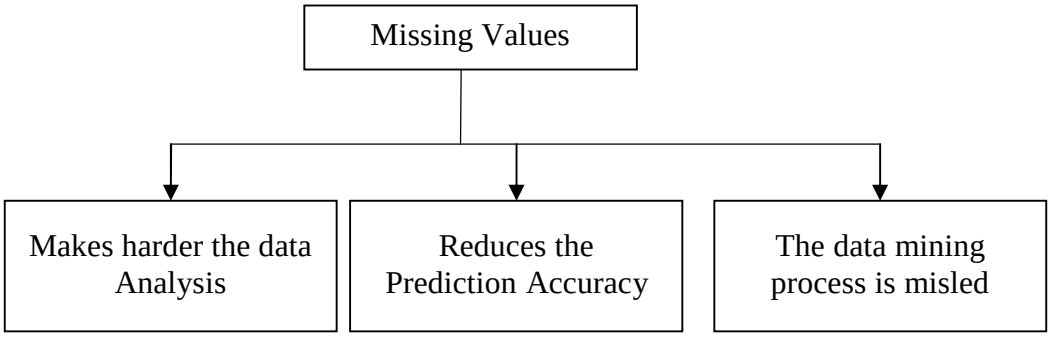


Figure 5.2: Missing value problems

Missing values are categorized [**Swa, 16**] based on the position which means all missing values are not harmful. The categorization of missing values is depicted in figure 5.3.

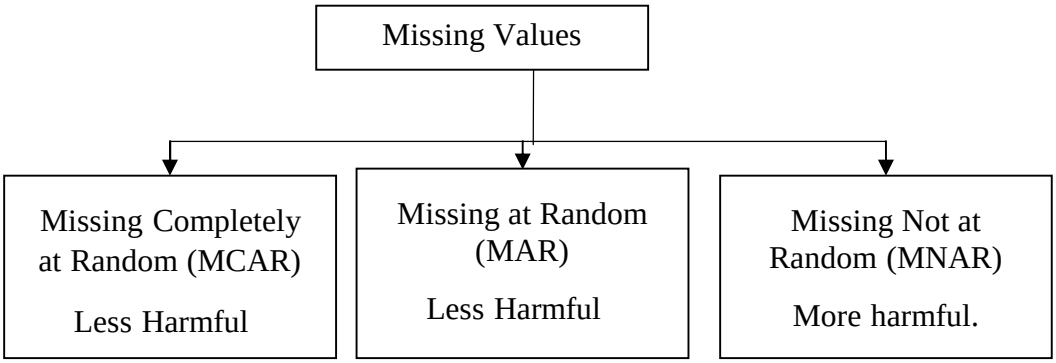


Figure 5.3: Categories of Missing values

MNAR type missing data does not exist in the proposed environment but, MAR and MCAR exist in the collected dataset. There is a difference between traditional missing values in the IoT environment and missing values in the mutual test environment. In a mutual test environment, parent sensor plays a vital role in solving missing values of child sensors. If the parent has missing values, then the child sensor plays a vital role to solve it. A detailed description of the missing values [**Arm, 17**] is discussed in table 5.2.

Table 5.2: Detailed description of missing values[Pap, 18]

<table>
<tr><th>Details</th><th>MAR</th><th>MCAR</th><th>MNAR</th></tr>
<tr><td>Name</td><td>Missing at Random</td><td>Missing Completely at Random</td><td>Missing Not at Random</td></tr>
<tr><td>Meaning</td><td>Data missing in the collected data and not in the target class.</td><td>Data missing in the rest of the target class or variables but not affecting the target class.</td><td>Data missing in the target class or variable but affecting the target class.</td></tr>
<tr><td>Method</td><td>Most of the Time Deletion is the recommended method. But the position of the value decides the method selection.</td><td>Most of the Time Deletion is the recommended method. But the position of the value decides the method selection.</td><td>The deletion method is not applicable for this Type. Interpolation or imputation methods are used based on the nature of the data.</td></tr>
<tr><td>Example</td><td>
<table>
<tr><th>T</th><th>R</th><th>H</th><th>S</th></tr>
<tr><td>x</td><td>350</td><td>40</td><td>Rain</td></tr>
<tr><td>38</td><td>x</td><td>70</td><td>Hot</td></tr>
<tr><td>12</td><td>880</td><td>x</td><td>Avg</td></tr>
</table>
T => Temperature
R => Rain
H => Humidity
S => Status of Water (Target class)
X =>Missing Values
R is a target variable in row1.
T, H are the target variables in row 2.
T is a target variable in row3.
</td><td>
<table>
<tr><th>T</th><th>R</th><th>H</th><th>S</th></tr>
<tr><td>x</td><td>350</td><td>x</td><td>Rain</td></tr>
<tr><td>38</td><td>x</td><td>70</td><td>Hot</td></tr>
<tr><td>12</td><td>x</td><td>x</td><td>Avg</td></tr>
</table>
T => Temperature
R => Rain
H => Humidity
S => Status of Water (Target Class)
X => Missing Values
R is a target variable in row1.
T, H are the target variables in row 2.
T is a target variable in row3.
</td><td>
<table>
<tr><th>T</th><th>R</th><th>H</th><th>S</th></tr>
<tr><td>15</td><td>x</td><td>X</td><td>Rain</td></tr>
<tr><td>38</td><td>880</td><td>X</td><td>Hot</td></tr>
<tr><td>x</td><td>x</td><td>50</td><td>Avg</td></tr>
</table>
T => Temperature
R => Rain
H => Humidity
S => Status of Water (Target Class)
X => Missing Values
R is a target variable in row1.
T, H are the target variables in row 2.
T is a target variable in row3.
</td></tr>
<tr><td>Problem</td><td>Harmless</td><td>Harmless</td><td>Harmful</td></tr>
</table>

5.2. Background Study

5.2.1 Missing Value Handling Techniques

Existing Missing values handling techniques are classified into three types and are listed in figure 5.4. Deletion method, imputation technique, statistical and prediction models are various techniques to handle missing values [**Lee, 19**]. The deletion method eliminates the data in three different manner, by using list wise,

pairwise and dropping variables [**Cur, 19**]. List wise deletion deletes the entire record from the data [**Hos, 20**]. The main problem of list wise deletion was data loss. So, pairwise deletion was introduced. In a pairwise deletion, particular variable was deleted [**Ran, 21**]. But if the deleted variable has variable dependency, then this method was not applicable. The variable dropping technique was introduced to

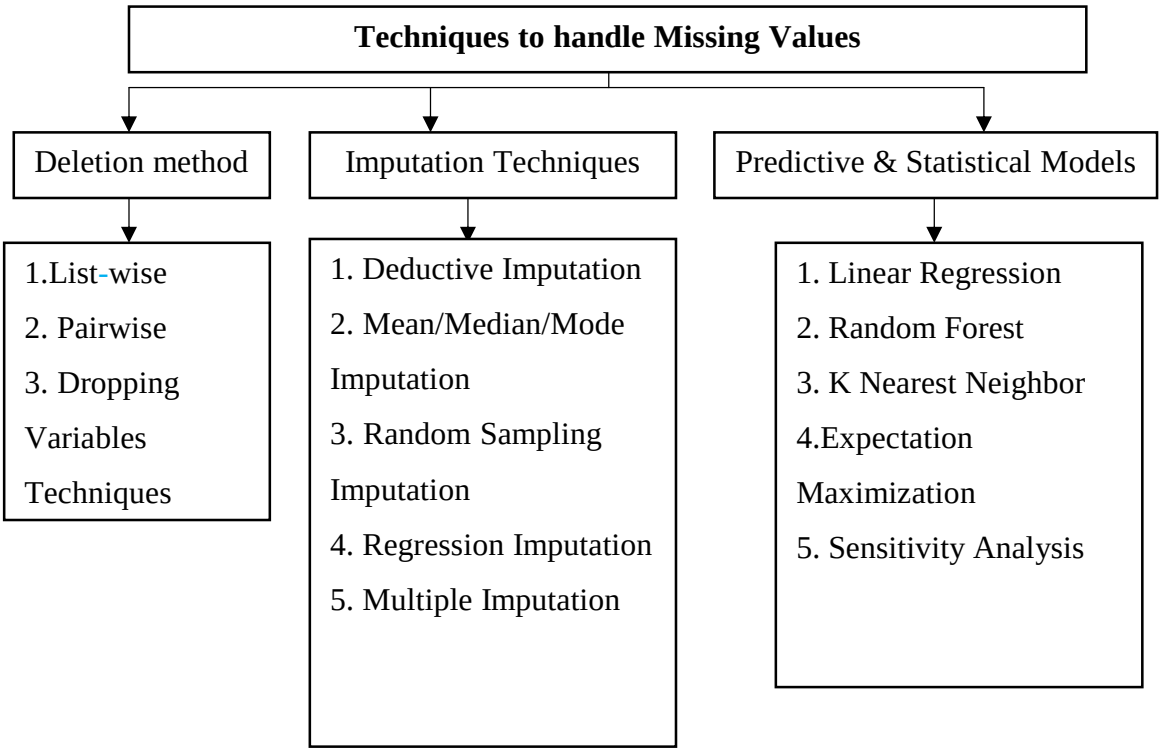


Figure 5.4: Techniques to handle Missing

overcome the drawback of pairwise deletion. This technique was used to select or to drop the dependent and independent variables. The main disadvantage of this method is the additional support which is needed to identify the variable dependency [**Cha, 21**].

Comparisons between various deletion methods are listed in table 5.3. The deletion method is suitable for MAR and MCAR values. The proposed environment has sensor errors in the parent sensors and collected data has a variable dependency. So the deletion method is not applicable to the collected data.

Table 5.3: Comparison of deletion methods

T	R	W	Status	List wise	Pairwise	Dropping Variable
20	380	NaN	Rain	Delete entire W	Delete W (Pair Value)	Select T, R
NaN	400	2536	Heavy rain	Delete entire T	Delete T (Pair Value)	Select R, W
T is Temperature Sensor Value (°C) R is Rain Sensor Value (Sensitivity Value) W is Anemometer (Wind sensor value) Feet Per Minute (FPM) T, R, W are all dependent variables.						

Imputation techniques are used to replace the missing values by using a data matrix. Statistical methods and other mathematical techniques are adopted for matrix creation. There are many methods in the imputation techniques, and figure 5.7 lists some of the popular methods. Generally, the imputation method is selected based on the nature and behavior of the data. Table 5.4 discussed the types of imputation techniques with their advantages and disadvantages while handling traditional and mutual test environment data.

Table 5.4: Imputation techniques for traditional IoT and mutual test IoT environments

Method	**Working**	**Traditional IoT Data**	**Mutual test IoT Data**
Deductive Imputation [**Lin, 20**]	This technique uses logic or statistical measures to find the relationship between the variables for replacing the missing values.	This is the simplest and best method because the IoT variable has dependency and relation with other variables.	Variable dependency and relationship between variables are not constant. So, this method is not suitable. For example, in the rainy season, target class depends on the rain sensor values, but in the sunny season, the rain sensor values
Mean/Median/Mode Imputation [**Had, 20**]	This method uses the mean or median or mode value to replace the missing value, error value and inconsistent value.	This method is not recommended for traditional IoT applications because this method suffers when the variable has relationship between them. The distance between the attribute values are not common, hence this method is not recommended	This method is suitable for mutual test environment because of the single variable (parent sensor) dependency on the mutual test environment. DaRoN technique has used central tendency values to handle noise. The distance between the attribute values is common thus this method is used.
Random Sampling Imputation [**Wij, 20**]	This method uses a random sample of the variable for filling the missing values.	This method uses random sampling based on the seeder value. But this method does not affect the original observations hence this method is applicable.	If a missing value exists only in the child sensor, then this method is the best solution. Unfortunately, in some places, sensor error occurs in the parent sensor, so this method is not applicable.

Regression Imputation [**Ver, 20**]	The regression model is used to fill the missing values based on the variable dependency.	This model used a regression model to predict and replace missing values. Hence, it is recommended for traditional IoT data.	This method is also recommended if the parent sensor does not have a sensor error.
Multiple Imputation [**Huq, 18**]	This method uses multiple imputations to fill the missing values which mean the best values are selected from the imputation set for replacement.	This method provides a set of values for replacement and the best value is selected. Hence, it is recommended.	Huge training time and over fitting problems are the two reasons for not selecting this method. Otherwise, this method is applicable.

Statistical methods and prediction models are used to fill the missing values. Based on the variable values, methods are selected [**Tan, 17**]. Generally, statistical methods are working based on distributions which take a lot of time. Hence, it's not suitable for the mutual test IoT environment. Popular statistical methods are listed in figure 5.4.

5.3. RELATED WORKS

Tao et al., [**Tao, 19**] proposed Incremental Space Time-based Model (ISTM) for missing value imputation. The author initially converted IoT stream data into a data matrix. The ISTM was used to reduce the data matrix conversion time. Neighbor matrix values were used to replace the missing value. In case if both the neighbors have missing values then this method will not be applicable.

Du et al., [**Du, 20**] predicted the base station traffic by using the Extreme gradient boosting-long–short-term memory (XGBoost-LSTM). Min, Max and Mean

values were used to replace the error values. The author claimed error values as missing values. IoT traffic dataset was used by the model and for prediction, XGBoostori, LSTMori, Co_Modelori, and Co_Modelini were used by the XGBoost-LSTM model. Training time increased a lot while working with outliers. So, the author concluded by stating that, if sensor errors were present in the dataset, then this method will not be applicable.

5.4. NEED FOR RESEARCH

Missing values occur in the form of Sensor errors in the collected data. In some places, the Parent sensor is also affected by the sensor errors. Figure 5.5 shows the sensor error details in the collected data.

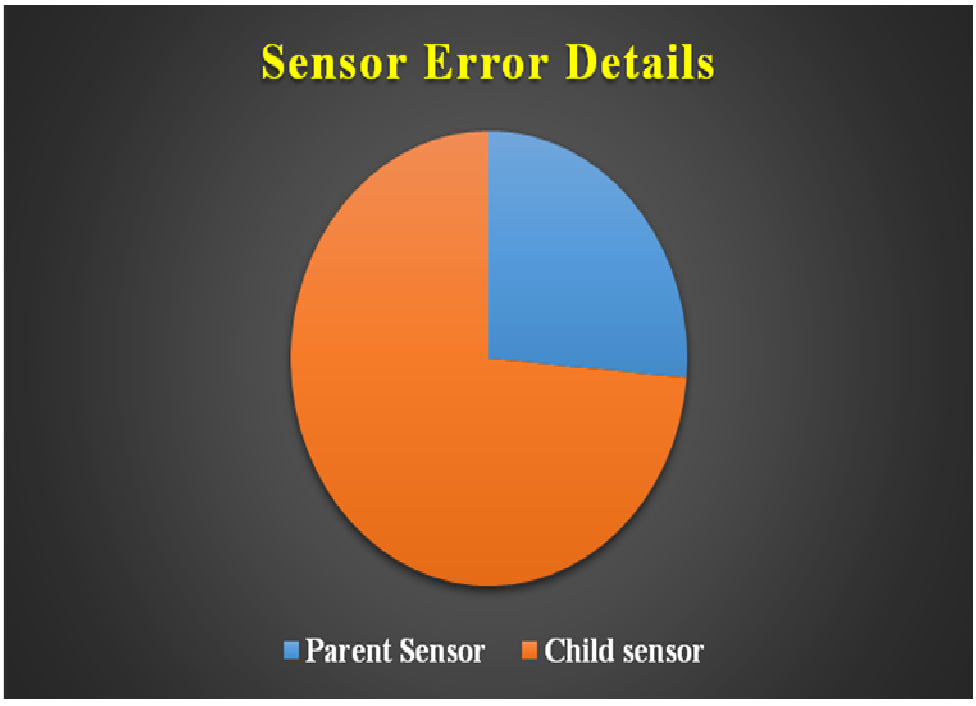


Figure 5.5: Sensor Error Details

Figure 5.6 shows the harmless types of missing values occur in the data. But, this has to be treated because the parent sensor is affected by missing values.

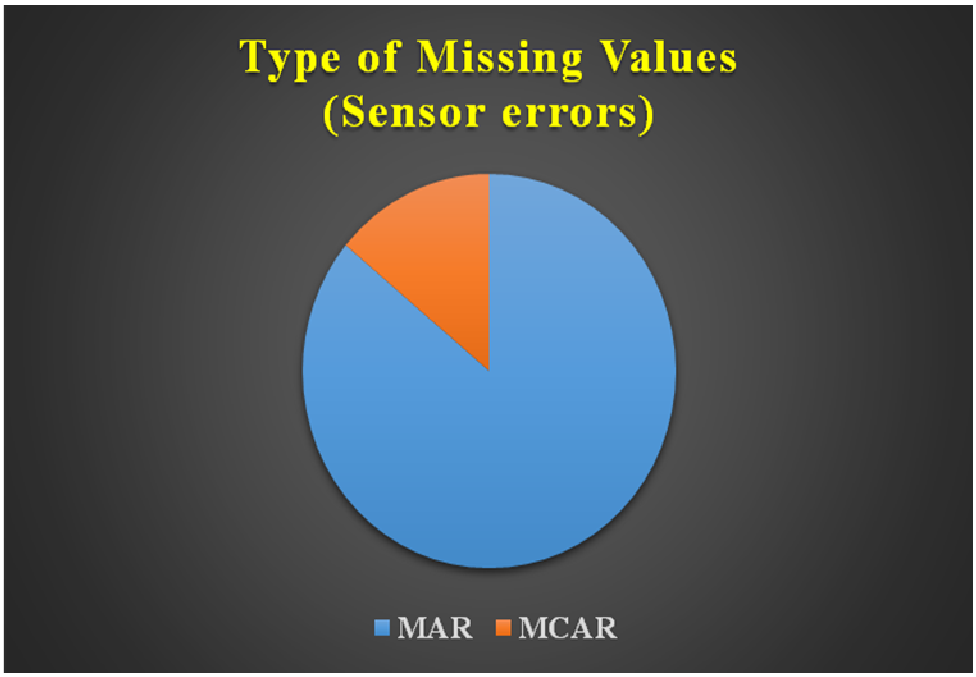


Figure 5.6: Type of Missing values (Sensor errors)

Figure 5.7 shows that the existing methods need to be improved for handling this data. So research is needed to handle these types of problems. TANOS is proposed to handle these sensor errors with better performance.

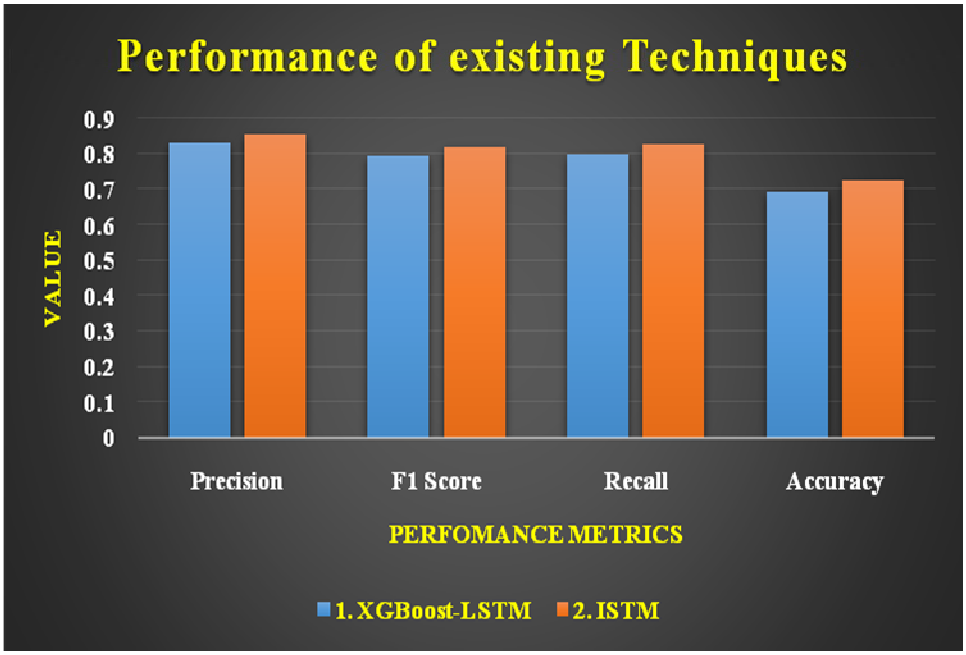


Figure 5.7: Performance of Existing Techniques

5.5. OBJECTIVES

The chapter aims to detect and remove sensor errors in the dataset to improve the accuracy in decision making. Sensor errors affect the quality of the dataset and therefore requires techniques to handle both parent and child sensor errors. So, the objectives of the proposed work are,

- To handle sensor errors by using the imputation technique.
- To handle sensor errors from both neighbors i.e., child sensors and parent sensors.

5.6. METHODOLOGY DIAGRAM OF THE TANOS TECHNIQUE

Generally, a mutual test environment partially removes missing values and outliers by using parent sensor. But the collected data has missing values in the form of sensor errors and this error exists in some elements of parent sensor. This error occurs due to power failure or connection failure or sensor failure. The proposed TANOS technique is capable of handling sensor error in an efficient way than other existing techniques. The methodology of the proposed TANOS technique is shown in figure 5.8. NaN is the only sensor error that exists in the collected data. NaN sensor error is identified using the keyword NaN. Initially, the error checker checks the error in the child sensor if the sensor error is found, then that particular child value will be replaced by the assigned neighbor value. If both have sensor error then it will be replaced by the parent sensor value. After removing all errors in the child sensors, the error checks to the parent sensor.

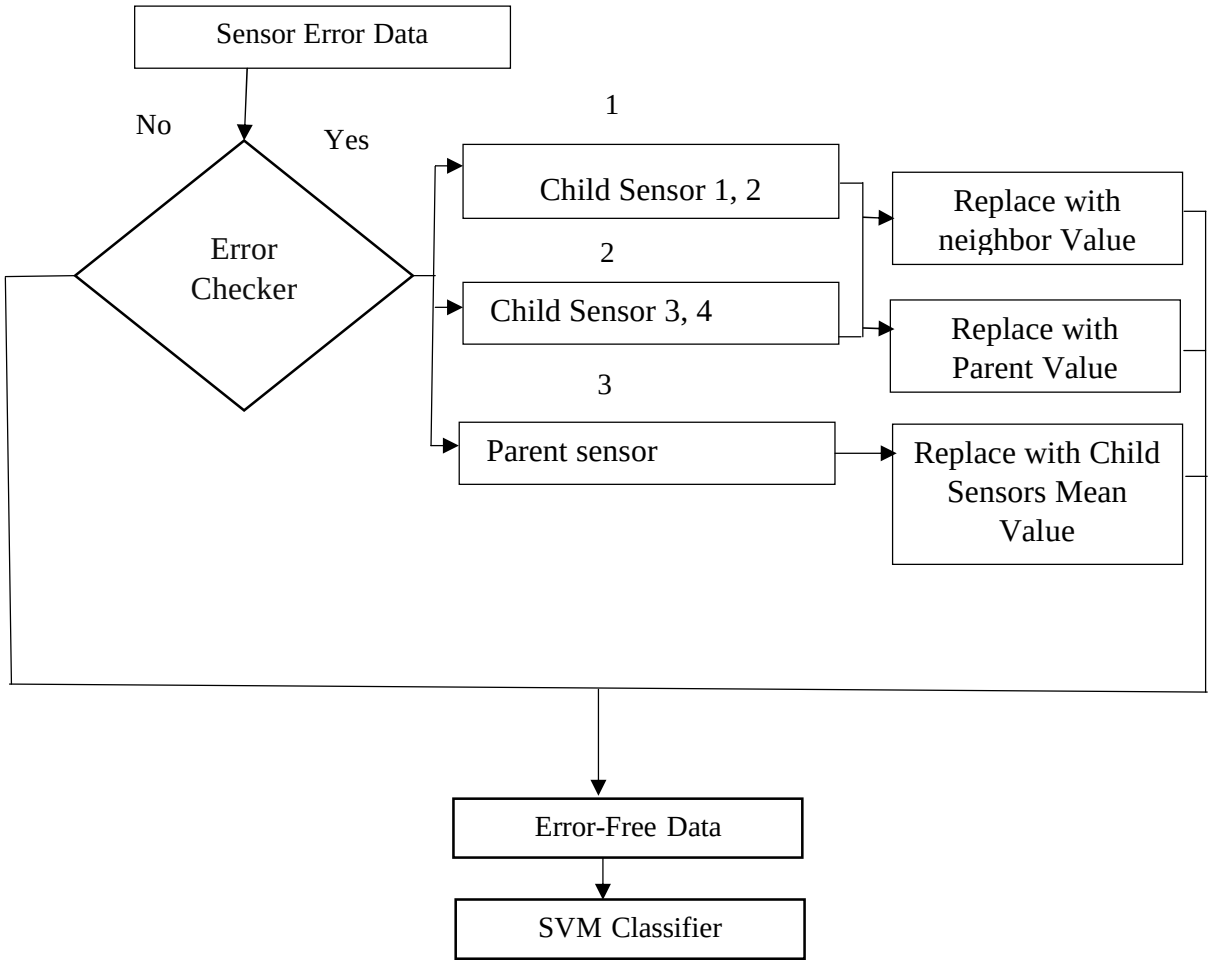


Figure 5.8: Workflow of TANOS technique

In some places, the parent sensor has sensor errors that are replaced by child sensors mean value. This process is continued for all 25 sensors used in the proposed Mutual test IoT environment. Finally, errors are removed from collected data and the data will be error-free. TANOS removes the matrix conversion problem by using mean values and handles the situation that is if both neighbours have missing

5.6.1. Working Procedure of TANOS technique

NaN error exists in the proposed environment. Initially, error positions and missing value types are verified.

Initially, Neighbors are assigned for each sensor child sensor 1 is the neighbor of child sensor 2. If child sensor 1 has a sensor error, then replace it by using child sensor 2 & vice versa. This process is similar to child sensors 3 and 4. Figure 5.9 shows that the process of assigning neighbors.

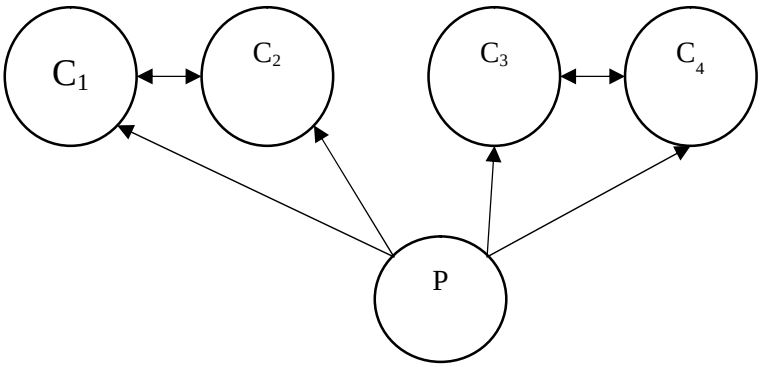


Figure 5.9: Assigning the neighbor

Where,

C_1, C_2, C_3, C_4 = Child sensors.

P = Parent sensor

In the proposed mutual test environment sensor 1^{st}, 2^{nd}, 3^{rd} and 4^{th} are child sensors and 5^{th} sensor is parent sensor, for example t_1, t_2, t_3, t_4 are child sensors and t_5 is the parent sensor

(t = temperature sensor).

$t_1 = \{t_2, t_5\}$, $t_2 = \{t_1, t_5\}$, $t_3 = \{t_4, t_5\}$, $t_4 = \{t_3, t_5\}$, $t_5 = \{t_1, t_2, t_3, t_4, t_5\}$

t_1 has two neighborst_2, $t_5$and t_2 hast_1t_5 as neighbors similar for $t_3$and t_4 respectively.

t_5 is the common neighbor for all child sensors.

If t_1 has a sensor error, then replace it by using the t_2 value. In case if t_1 and t_2 have sensor error, replace it by using the $_{t5}$ (parent sensor) value. Similar for t_3 and t_4. This step handles sensor errors that occur in child sensors.

If the parent sensor has an error value, it will be replaced by the mean value of child sensors.

$$t_5 = \frac{t_1 + t_2 + t_3 + t_4}{4}$$

TANOS uses mean imputation because the previous noise removal technique DaRoN used the central tendency measures. So, that the deviation problem caused by the mean imputation method is avoided. Thus, TANOS handles sensor errors efficiently and enhances the accuracy of dataset making it easier to predict and make decisions in an error free manner. The steps involved in the TANOS technique is discussed below,

5.6.2. Steps for TANOS technique

Step 1: Load data D

Step 2: Initialize all attributes namely T, H, W, R and S

Step 3: Define the Parent sensors and child sensors to replace the NaN values

Step 4: Assign neighbors for all sensors

Step 5: If child sensor value is NaN then the value of its parent sensor is replaced likewise the parent sensor is NaN then it is replaced by its child sensor value.

Step 6: If the parent sensor t_5 only NaN then it is replace by the mean value of t_1, t_2, t_3 and t_4

Step 7: If all sensor values are NaN then the entire row is deleted

Step 8: Missing values are replaced and the data is fine-tuned

5.6.3. TANOS Technique

Technique: TANOS

Input: Noise removed and balanced data with sensor error

Output: Error-free Data

Load Data set D ***// After cleaned by DaRoN***

Assign N (t_1) =t_2 ***// t_1 and t_2 are neighbors***

Assign N (t_3) = t_4 ***//t_3 and t_4 are neighbors***

Assign N (t_5) = t_1, t_2, t_3, t_4 ***//t_1, t_2, t_3, t_4, are neighbors of t_5***

//neighbors assignment is similar for all elements in R, S, W and H

If t_1 = "NaN" && t_2 = "NaN" && t_3 = "NaN" && t_4 = "NaN" && t_5 = "NaN"

Remove entire row

else if t_1= " NaN" ***// similar for all elements in R, W, S, H***

replace with t_2 value ***// NaN denotes sensor Error NaN***

else if t_2= " NaN" ***// similar for all elements in R, W, S, H***

replace with t_1 value

else if t_1= " NaN" and t_2 = "NaN"

replace with t_5 value ***// similar for t_3, t_4 and all elements in R, W, S, H***

else if t_5 = "NaN" ***//infrequent case***

replace with Mean Value of t_1, t_2, t_3, t_4 ***// similar for all elements in R, W, S, H***

else

Keep the original value ***//unchanged***

end if

end else

5.7. RESULTS AND DISCUSSIONS

In this research work, data are collected in a mutual test environment. So, they can't be compared to traditional methods because in mutual test environment sensor has priorities as parent and child. Figure 5.7 proved that the existing techniques are not suitable for handling the mutual test environment data. TANOS technique is applied to the SVM classifier to check the performance of the dataset after applying the proposed technique. Confusion matrix metrics (precision, recall, accuracy, f1 score) are used to evaluate the performance of TANOS. TANOS technique is compared with DaRoN technique to prove the improvement in the performance of the dataset. In figure 5.10 performance metrics chart, X-axis represents the performance metrics and the Y-axis represents the confusion matrix values.

TANOS achieved 97.03% precision whereas DaRoN achieved 96.05% precision. It has positively predicted values from the collected data. F1 score is calculated by using precision and recall. TANOS achieved 92.26% whereas DaRoN achieved 90.27% F1 score. Recall is the detection possibility of the classifier. TANOS achieved 87.14% whereas DaRoN achieved 85.14% recall. Accuracy is the fraction between correct prediction and total prediction. TANOS achieved 85.18% whereas DaRoN achieved 84.18% accuracy. The result show that TANOS technique outperforms DaRoN technique. The performance of the SVM classifier is enhanced. Figure 5. 10

shows that the performance of DaRoN and TANOS in terms of all confusion matrix metrics.

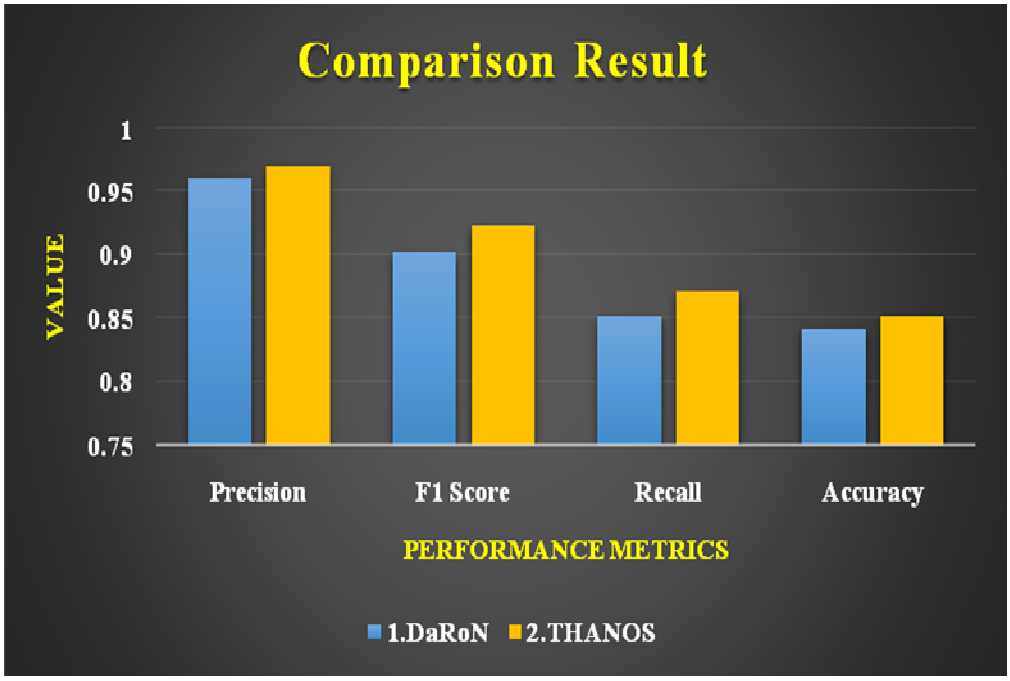


Figure 5.10: Comparison Result

In the above figure 5.10 proposed TANOS technique outperforms DaRoN technique in all metrics. DaRoN removes noisy data but does not handle the sensor errors. But TANOS handles the sensor errors also. After the DaRoN and the TANOS technique are applied to the collected IoT irrigation dataset, all noises, including sensor errors (missing values) are removed. Finally the dataset becomes error-free.

Error handling is the reason for the raise in classifier accuracy and precision. DaRoN replaced the noise values with the mean values so that all elements in the target class are filled except the sensor errors. This leads to a balance in data partially. But after the TANOS technique is applied, the data set is fully balanced because TANOS removed sensor errors that affect the target class. Now the data set has equal number of elements in the target class.

5.8. FINDINGS AND INTERPRETATIONS

This section explains the experimental results of the DaRoN technique and the proposed technique TANOS. It explicates the strength of the TANOS technique and how TANOS achieves better results compared with DaRoN technique. It also justifies how the TANOS technique is better than the DaRoN technique.

The DaRoN technique was used as the central tendency measure to remove noise in the mutual test IoT environment data. But DaRoN has not focused on sensor errors. So TANOS technique is proposed to handle sensor errors.

In this research, the DaRoN technique and TANOS technique are applied to the SVM classifier to evaluate the performance of both techniques. TANOS achieved better performance than DaRoN in all performance metrics. So TANOS improved the SVM classifier performance better than DaRoN. Table 5.5 shows the comparison of DaRoN and TANOS techniques.

Table 5.5: Comparison of DaRoN and TANOS

Technique Name	Performance Metrics			
	Precision	**F1 Score**	**Recall**	**Accuracy**
DaRoN	0.9605	0.9027	0.8514	0.8418
TANOS	0.9703	0.9226	0.8714	0.8518

5.9. CHAPTER SUMMARY

Traditional missing value handling techniques have not focused on the problem when both neighbors have missing values. There is a possibility of the occurrence of over fitting problem while using existing methods and these methods are not capable of handling data collected under a mutual test environment which have missing values in the form of sensor errors. TANOS technique has overcome this hurdle and handled

all missing values. All sensor errors are in the form of Missing Completely At Random (MCAR) and Missing At Random (MAR). The position of sensor errors sometimes affects the target class so that imputation techniques are used otherwise, deletion technique is the preferable method. The proposed TANOS achieved better performance than the DaRoN technique. TANOS provided good results in terms of all metrics of the confusion matrix. TANOS technique handled the neighbor value sensor error problem and removed the over fitting problem. Further, to increase accuracy, these noise-free data are applied to feature selection techniques to eliminate the irrelevant features based on the environmental season conditions in the next chapter.

Chapter – 6

MESIA: enseMble filtEr based feature Selection for IoT Agriculture Data

Chapter - 6

MESIA: enseMble filtEr based feature Selection for IoT Agriculture Data

6.1. INTRODUCTION

Feature selection is a preprocessing method used to reduce feature space and to select the best subset from the collected dataset features. Feature selection provides a successful reduction in dimensionality, noisy data and features removal, improves learning algorithm accuracy (Machine/ Deep Learning), also reduces the model building time and decision making efficiency [**Sun, 20a**]. The benefits of feature selection are summarized in figure 6.1.

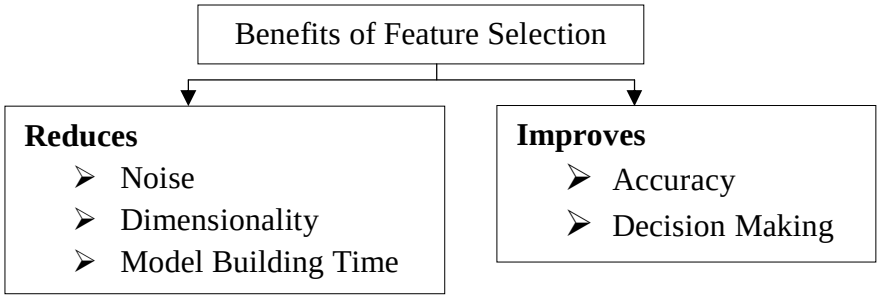


Figure 6.1: Benefits of Feature Selection

IoT sensors which are used in this proposed mutual test environment make the data collection process much easier, but the high dimensionality (number of rows) of the collected data and growing speed of data directly affects the machine learning process. There are three methods to handle features. They are: feature selection, feature extraction and dimensionality reduction. Feature handling methods are discussed in Table 6.1. Based on the table 6.1, feature selection is selected for the proposed mutual test environment [**Ait, 19**]. Feature selection is the best choice against data with huge noise and features that have relationships between them.

Table: 6.1: Difference between Feature Selection, Feature Extraction and Dimensionality Reduction

Category	Feature Selection	Feature Extraction	Dimensionality Reduction
Other Name	Attribute / Variable Selection	Dimensionality Extraction	Dimension Reduction
Definition	Selecting the required features from the collected Data.	Creating a feature by combining existing features	Transforming the data dimensionality from high to low
Process	Selection	Creation	Elimination
Applications [**Mia, 16**]	Mammographic Image Analysis, Criminal Behavior Modeling, Genomic Data Analysis, Plant Monitoring, Mechanical Integrity Assessment.	Bag of Words (Text Modeling), Image Processing, and Auto-encoders	Image Processing, Map and Navigation System
Methods [**Sha, 20**]	Filter, Wrapper, and Embedded	All Neural Networks based methods, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA)	Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Generalized Discriminant Analysis (GDA)
Role in Data Analytics	Data Cleaning	Data Cleaning and Data Reduction	Data Reduction
Applicable for	Data with a huge noise and features in relation	High Dimensional and data having more features	High Dimensional Data

Plant monitoring is one of the applications of feature selection which is used in the proposed environment. Dimensionality reduction and feature extraction methods are preferred while using image data [**Tad, 19**], [**Fer, 14**]. The feature selection methods used in the collected IoT irrigation data reduces the features, minimizes the processing time and increases classifier accuracy. Feature selection methods are classified into three types. They are filter, wrapper and embedded methods. Types of feature selection methods are shown in Figure 6.2.

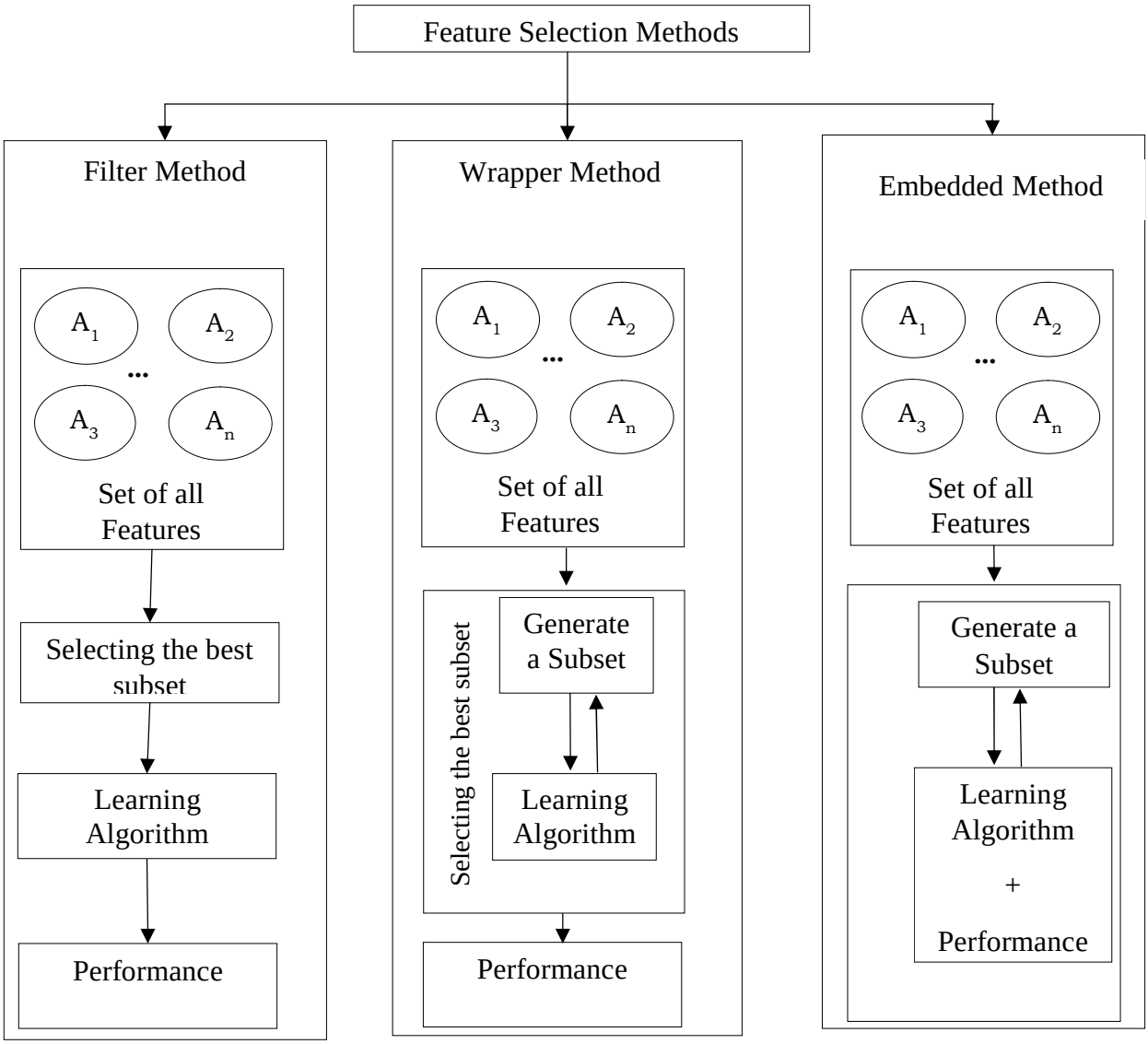


Figure 6.2: Types of Feature Selection methods

The filter method is the fastest feature selection method which is classifier independent, but it requires additional techniques to detect the variable dependency [**Zho, 19**]. The wrapper method is also a popular feature selection method but suffers with over fitting and low fitting problems [**Wah, 18**]. Filter methods are classified into two types univariate and multivariate. Univariate handles each feature individually whereas multivariate handles features based on the relationship between them [**Tsa, 19**], [**Sun, 18**].

The wrapper method is classified into two categories. They are deterministic wrappers and randomized wrappers [**Zen, 15**]. A deterministic wrapper results in an over fitting problem at a low level whereas a randomized wrapper has a huge risk of over fitting problems [**Aic, 19**]. Both wrapper methods [**Mao, 19**] are dependent on classifiers and take high computational time (CPU Time).The embedded method is the combination of filter and wrapper. This method is designed to remove the problems caused by wrapper and filter methods. Embedded method consists of two parts namely, filter and wrapper. Among the two parts of the embedded method, the filter part works well at all times. But, the wrapper part suffers with classifier dependency and fitting problems. Based on the features of the proposed IoT environment, the filter method is selected for the proposed work. Types of feature selection methods' are explained in Figure 6.3.

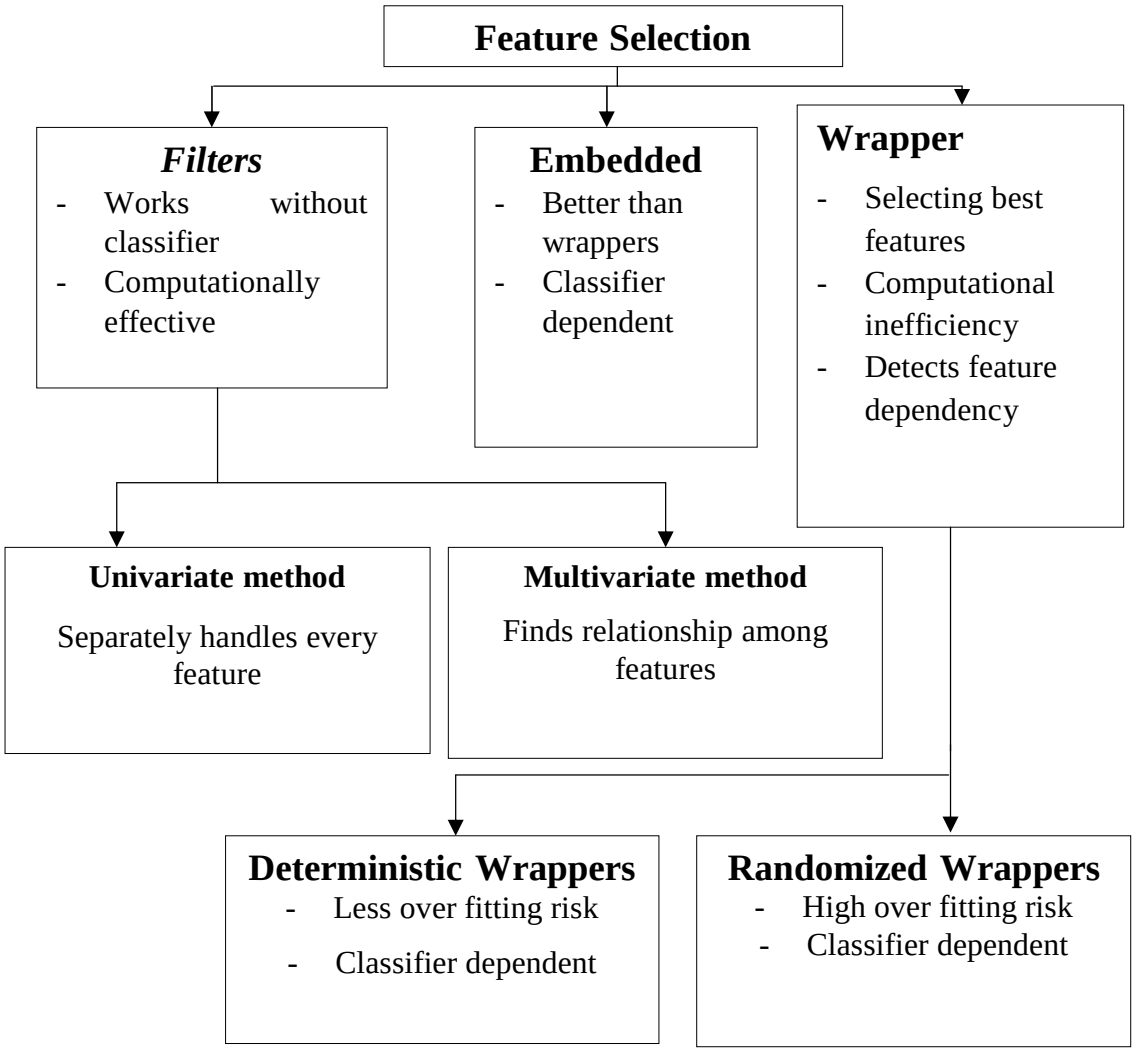


Figure 6.3: Classification of Feature Selection methods

6.2. BACKGROUND STUDY

6.2.1. Filter Method

Filter methods are fast in nature because they do not depend on the classifier. The only problem with the filter method is that, it requires additional support to identify the variable dependency, but variable dependency can be handled easily with the help of supporting techniques. Generally, supporting techniques are selected based on the nature of the data [**Luf, 21**]. Table 6.2 shows the example of variable dependency.

Table 6.2: Variable Dependency [**Ayv, 21**]

S.No.	**Temperature Sensor (T) °C**	**Rain Sensor (R) Sensitivity Value**	**Wind Sensor (W) Feet Per Minute (FPM)**	**Status**
1	37	100	300	Dry
2	27	356	1500	Wet

Threshold Values	**Values**	**Status**
	<350	Raining
	350 to 480	Heavy Rain
Rain Sensor (R)	480 to 640	Drizzling
	640 to 880	Rain Warning
	> 880	No Rain
Temperature Sensor (T)	>38^{o}C	Too Hot
	>34^{o}C to >38^{o}C	Hot
	>25^{o}C to >34^{o}C	Average
	<25^{o}C	Low
Wind Sensor (W)	<3432	Danger
	2201 to 3432	Average
	1761 to 2200	Low

In this table, the 1st row field status is determined based on the temperature value. So the rain sensor and wind sensor values are irrelevant because the temperature is too hot. In row 2nd temperature value is irrelevant because there is rain. In table 6.2, variable dependencies are identified by the season based conditions. For example, if the temperature sensor value is high and wind and rain sensors values are low, the status is updated as "**Dry"**. If the temperature sensor value is low and wind and rain sensors values are high, the status is updated as "**Wet"**. Figure 6.4 shows the supporting factors that can be considered for filtering technique.

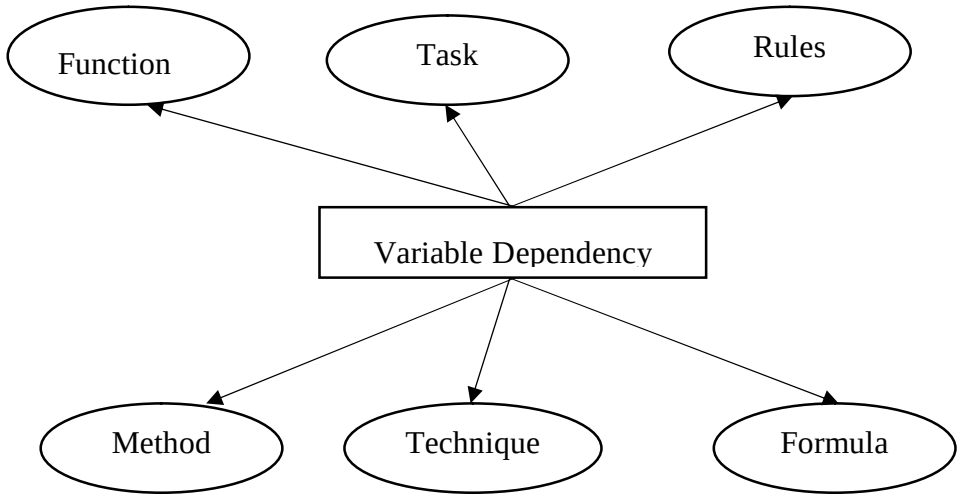


Figure 6.4: Supporting factors for variable dependency

The filter technique can be combined easily with other techniques because it is classifier independent. This independent nature of the filter techniques make it more computationally effective. Table 6.3 shows the existing filter techniques and their applications. Among these, statistical filter is used for multivariate filtering and user-defined filter is used for univariate filtering in the proposed MESIA technique.

Table 6.3: Existing filter techniques

Type of filter	Best Result	Application
Wavelet Transform Filter [**Dey, 21**]	Image and Signal Data	Signal processing
Kalman filter [**Vad, 21**]	Incomplete and chemical data	Corrupted data recovery and image processing
Statistical Methods [**Dev, 20**]	Features with relations	All smart applications use numerical data.
User-Defined [**Gio, 20**]	Based on the nature of data	Suitable for all applications.

6.3. RELATED WORKS

Aurora et al., [**Aur, 19**] proposed a feature selection methodology to handle time-series data. This methodology transforms the data into structured and standard form by using a Support Vector Machine (SVM) machine learning algorithm. The author used four types of feature selection methods. Univariate filter, multivariate filter, wrapper and multivariate wrapper which were used to select the best features. Multivariate wrapper provided better result among the feature selection methods. 10 fold validations were the feature selection methods used in this methodology to remove the over fitting problem caused by the multivariate wrapper, but this methodology suffered low fitting problem.

Egea et al., [**Ege, 17**] presented a strategy to select the best features in the Industrial Internet of Things (IIoT) data by using Fast Based Correlation Feature Selection (FBCF). Redundant features were handled by FBCF. Decision tree, Support Vector Machine (SVM), and Logistic Regression classifiers were used in this strategy. But, FBCF increased the work execution time.

Mohtashami et al., [**Moh, 19**] proposed a hybrid filter-based feature selection method to classify microarray (gene data) data. Rough sets, weighted rough set, fuzzy

rough set and hesitant fuzzy techniques were used to support the hybrid filter. Continuous features were handled by using approaches like weighted rough set dependency degree, information gain, and hesitant fuzzy approaches. The rough fuzzy theory was used to remove the redundant features. This method was fast but, could not handle negatively correlated features.

Table 6.4 shows the details of existing methods and their supporting factors.

Table 6.4: Existing Methods

Author	Technique	Supporting Factor
Aurora et al., [**Aur, 19**]	Filter and wrapper	Data conversion and 10 fold validation
Egea et al., [**Ege, 17**]	FBCF (Statistical method)	Regression
Mohtashami et al., [**Moh, 19**]	Hybrid filter	Fuzzy

6.4. NEED FOR RESEARCH

Mutual test data have continuous features and negatively correlated features which are not handled by existing techniques. The table 6.5 and figure 6.5 show the performances of the existing techniques against mutual test data.

Table 6.5: Comparison of existing feature selection methods

Authors Name	Methodology	Selected Features
Aurora et al., [**Aur, 19**]	Multivariate filter-based feature selection	29
Egea et al., [**Ege, 17**]	Fast Based Correlation Feature selection	25
Mohtashami et al., [**Moh, 19**]	Fuzzy rule-based	18

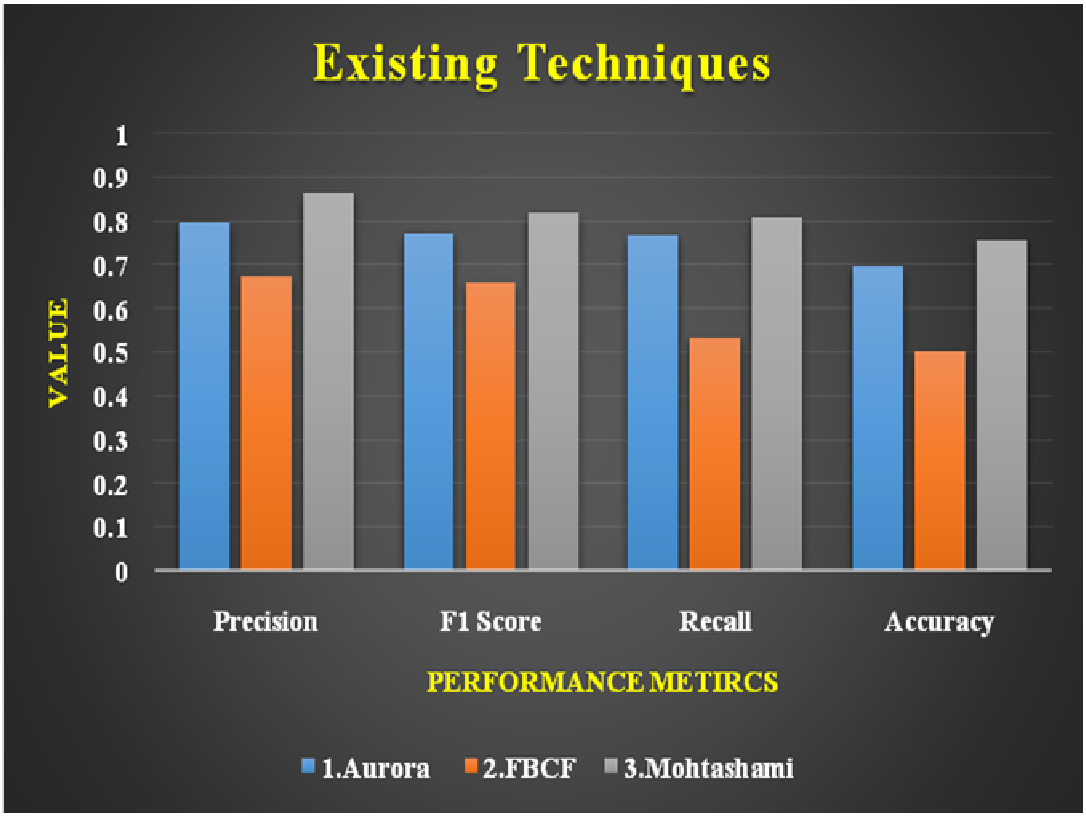


Figure 6.5: Existing Techniques performance

6.5. OBJECTIVES

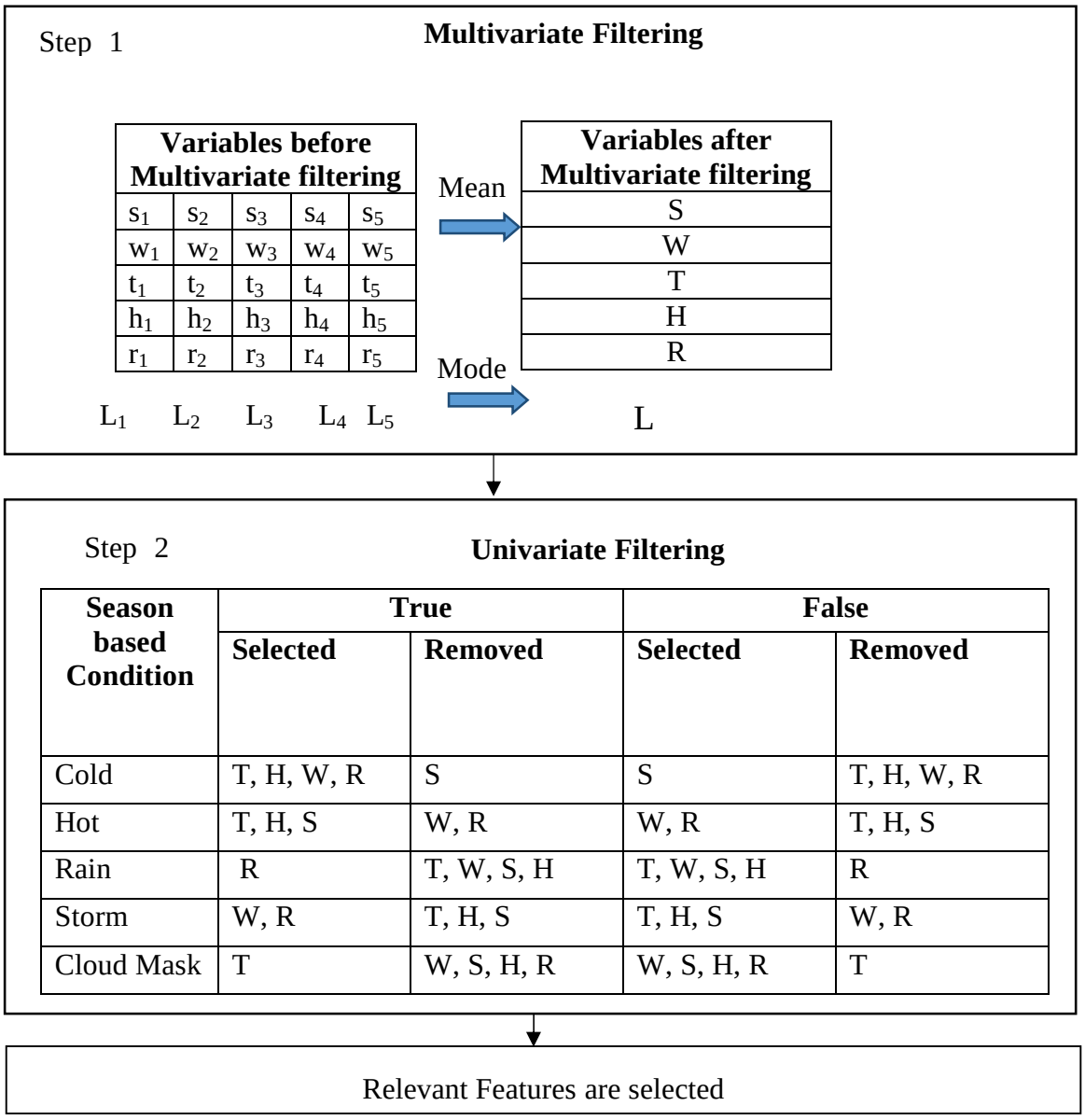
The chapter aims to detect and remove irrelevant features and continuous features in the collected dataset. The proposed MESIA technique also handles both positively and negatively correlated features. Feature selection increases the quality of the classifier and therefore requires techniques for all types of features. Thus, the objectives of the proposed work are,

- To remove irrelevant features by using the threshold values.
- To handle continuous features by using the Pearson correlation
- To handle both positively and negatively correlated features.

6.6. METHODOLOGY DIAGRAM

In the previous work, the DaRoN technique removed noisy data by using central tendency and the TANOS technique handled sensor errors by using the neighbour value, the mean value of child sensors, and parent sensor value. Now, the collected irrigation data is cleaned, but relevant and irrelevant features are not handled in the previous work. MESIA technique is proposed to select relevant features and to remove the irrelevant features simultaneously. MESIA technique uses ensemble filters to perform feature selection. A multivariate filter is used to combine the collected features by using mean values. DaRoN and TANOS technique uses mean values for handling sensor errors. So MESIA technique also uses mean values for filtering. MESIA technique methodology diagram is given in figure 6.6.

After multivariate filtering, univariate filtering is performed by using season based threshold values. These threshold values are assigned based on the sensor values and its status. The supporting factor for the univariate filter is the threshold values and the factors are shown in table 6.2. Univariate filter handles negatively and positively correlated features by using Pearson correlation. Pearson correlation is the popular correlation method used to find the relation between the data.



Season based Condition	True		False	
	Selected	Removed	Selected	Removed
Cold	T, H, W, R	S	S	T, H, W, R
Hot	T, H, S	W, R	W, R	T, H, S
Rain	R	T, W, S, H	T, W, S, H	R
Storm	W, R	T, H, S	T, H, S	W, R
Cloud Mask	T	W, S, H, R	W, S, H, R	T

Figure 6.6: Methodology of MESIA

6.6.1. Working of MESIA technique

Data are collected by using 5 sets of sensors (Soil Moisture sensor (S), Temperature Sensor (T), Humidity Sensor (H), Rain Sensor(R), and Wind speed sensor (W)) which are placed in 5 different locations. So, totally there are 25 sensors. T, S, H, W and R represents each sensor set and each has 5 elements (members) as listed below.

$T = \{t_1, t_2, t_3, t_4, t_5\}$

$S = \{s_1, s_2, s_3, s_4, s_5\}$

$H = \{h_1, h_2, h_3, h_4, h_5\}$

$R = \{r_1, r_2, r_3, r_4 r_5\}$

$W = \{w_1, w_2, w_3, w_4, w_5\}$

$L_1 = \{t_1, s_1, h_1, r_1, w_1\}$ similarly for L_2, L_3, L_4, L_5

Therefore, L can be written as L = {T, S, H, R, W}

There are 30 features (except for the timestamp and Target Class L). The timestamp feature is independent, but all other features are dependent on the time-stamp. So it remains unchanged during the filter process. Initially, a multivariate filter combines the sensors values by using mean values. In the previous DaRoN and TANOS techniques, most of the time parent sensor value is selected to replace the child value. This mean value is the nearest value or the same value while compared with the parent sensor value. So, the mean value represents actual data.

L_1, L_2, L_3, L_4, L_5 are Target class values concerning their location which are also common most of the time. The mode is calculated by the number of occurrences of a value and it represents L's status. For example L_1, L_2, L_3, L_4, L_5 = {wet, dry, wet, wet, wet} then L's status will be "*Wet*" based on Mode calculation. After the mode is calculated, the multivariate filter uses the Statistical method as a supporting factor to select relevant features. Finally, after Step 1 there are only 7 features.

In MESIA technique, banana plant irrigation data is collected. Generally, this plant requires a lot of water, but during the rainy season, watering has to be limited or stopped for the plant's health. So, the first condition is framed to handle the rainy season. The univariate filter uses season based threshold values and Pearson

correlation factor as supporting techniques to perform filtering. This condition uses a positive Pearson correlation to perform filtering.

During the cold season, the photosynthesis process is partially reduced so that the plant does not need water until the water level reduces in the soil. In this season, there is no correlation between the features, so threshold value is used to perform filtering.

During the sunny season, the banana plant needs continuous water supply until the water level reaches its max threshold. This season does not have correlation between features. So, threshold value is used for filtering.

During the storm season, the banana plant does not need a continuous water supply until the Wind speed level reaches its minimum threshold. This season has correlation between features so, threshold values and positive Pearson correlation are used for filtering.

Cloud mask is a special condition where most of the existing agricultural applications fail to handle it. In this season, the sun is hidden behind the clouds. During this time temperature sensor readings are low. Sometimes this occurs frequently in a day. In some cases, water is required and sometimes water is not required. This condition may prolong for several hours in a day. During this season water supply may be needed from time to time or not. So, this situation has to be treated specially. Negative Pearson correlation and threshold values are used to perform filtering. Table 6.6 shows the features selected by the univariate filtering.

Table 6.6: Selected features details

Selected feature(s) for the rainy season	**True**	**False**
	1 (R)	4 (T, S, H, W)
Selected feature(s) for the cold season	4 (T, W, H, R)	1 (S)
Selected feature(s) for the hot season	3(T, H, S)	2 (R, W)
Selected feature(s) for storm season	2 (W, R)	3 (T, H, S)
Selected feature(s) for cloud mask season	1(T)	4(H,W,S, R)

The steps involved in the MESIA technique is discussed below,

6.6.2. Steps for MESIA

Step 1: Load data D

Step 2: Initialize all attributes namely T, H, W, R and S

Step 3: Multivariate filtering is applied to remove unwanted data based on the seasonal condition

Step 4: For that, initially mean values are calculated for all sensors T, S, R, W and H.

Step 5: Find out mode value for class label (Water status of L_1, L_2, L_3, L_4 and L_5)

Step 6: After univariate filtering is applied to remove the unwanted data based on the seasonal condition.

Step 7: In univariate filtering, season based conditions are defined including rainy season, cold season, summer season, storm season and cloud mask season

Step 8: Minimum and maximum threshold value is defined for all attributes T, S, R, W and H using the mode value

Step 9: The unwanted data is removed by comparing the mean values of T, S, R, W and H with its threshold values

Step 10: Cleaned data is produced

6.6.3. MESIA technique

Input: Cleaned Data with all features

Output: Data with Limited features

Initialisation:

Load Data D **// After cleaned by TANOS technique**

Import statistics

Initialize Threshold value **// Refer table 6.2**

Compute $S_{Mean} = \frac{s_1 + s_2 + s_3 + s_4 + s_5}{5}$ **// S = Soil Moisture sensor values**

Compute $T_{Mean} = \frac{t_1 + t_2 + t_3 + t_4 + t_5}{5}$ **// T= Temperature Sensor Values**

Compute $W_{Mean} = \frac{w_1 + w_2 + w_3 + w_4 + w_5}{5}$ **// W = Wind Sensor Values**

Compute $H_{Mean} = \frac{h_1 + h_2 + h_3 + h_4 + h_5}{5}$ **// H = Humidity Sensor Values**

Compute $R_{Mean} = \frac{r_1 + r_2 + r_3 + r_4 + r_5}{5}$ **// R = Rain Sensor Values**

Compute st.Mode([L_1, L_2, L_3, L_4, L_5]) // **Frequent water status is selected by using Mode**

if ($R_{Mean} > R_{Min}$ and $R_{Mean} < R_{Max}$) **//Rainy Season**

Remove T, S, H, W until $R_{Mean} < R_{Min}$

Compare mean of (r_1,r_2,), (r_2,r_3,), (r_3,r_4,), (r_4,r_5,) with R_{Min} and R_{Max}

if (Mean > R_{Min} and Mean <R)

remove T, S, H, W

If ((R ρ T) && (R ρ H)) **//ρ= Correlation**

remove T, S, H, W **//Correlation found**

else

remove R **// correlation not found**

else if ($S_{Mean} < S_{Min}$ *and*) *and* $H_{Mean} > H_{Max}$ *and* $R_{Mean} > R_{Min}$ *and* $R_{Mean} < R_{Max}$ *and* $T_{Mean} > T_{Min}$) **// *Cold Season***

remove W

If ($W_{Mean} > W_{Max}$)

remove all features from T, S,H, R

else

remove T, H, W, R until $S_{Mean} < S_{Max}$

else if ($T_{Mean} > T_{Max}$ and $H_{Mean} > H_{Max}$ and $S_{Mean} < S_{Min}$) **// Hot Season**

remove R, W

else if ($H_{Mean} > H_{Min}$ *and* $S_{Mean} < S_{Min}$ *and* $T_{Mean} > T_{Min}$ *and* $R_{Mean} > R_{Min}$ *and* $R_{Mean} < R_{Max}$) **//Storm Season**

remove T, H

else if ($W \rho R$)

remove T, H, S **//Correlation found**

else

remove R **//Correlation not found**

else if ($T_{Mean} < T_{Min}$ *and* $H_{Mean} > H_{Min}$ *and* $S_{Mean} < S_{Max}$) **//Cloud Mask**

remove H, W, S, R

if ($T - \rho S$) **//-ρ Negative Correlation**

remove S **//Correlation found**

else

remove T **//Correlation not found**

end if

end if else

end else

end

By using the above technique it is proven that MESIA technique successfully handled the feature selection process. Figure 6.7 shows, how the MESIA technique handled feature selection. MESIA technique handles the continuous features by using the mean and the mode values. Correlated features are handled by using threshold values and positive and negative Pearson correlations. MESIA technique performed on both column and row wise filtering. Multivariate filter performed column filtering, and univariate filter performs both column and row wise filtering.

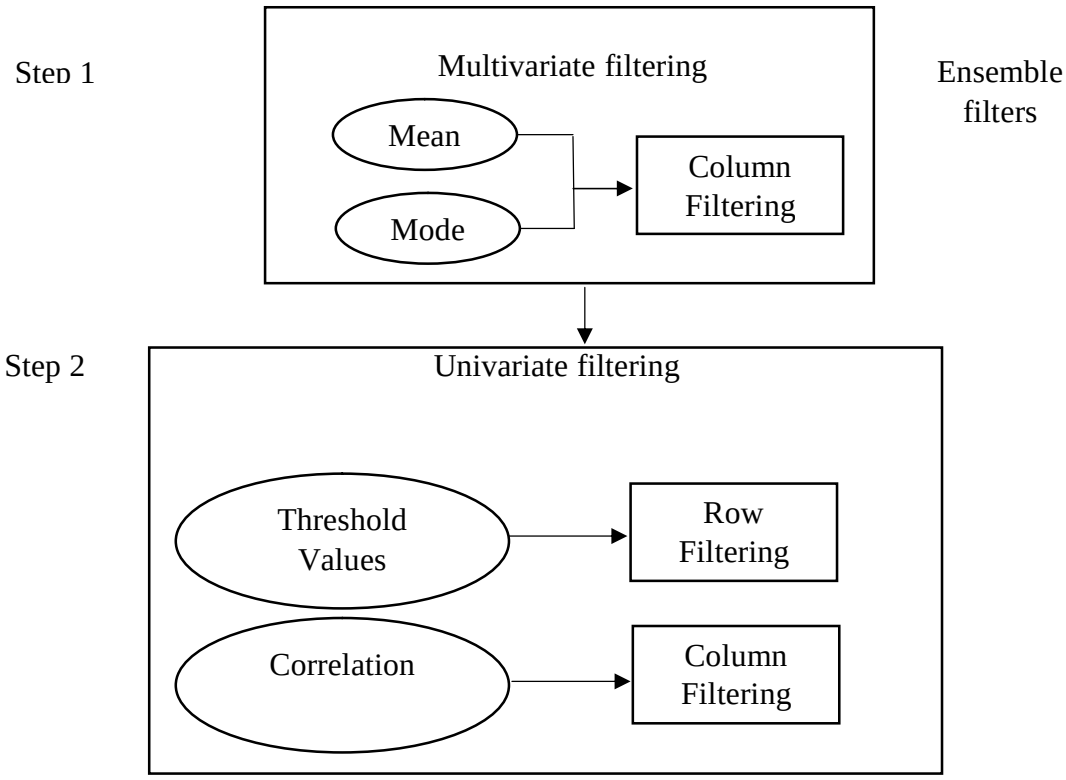


Figure 6.7: Features Selection by MESIA Technique

6.7. RESULTS AND DISCUSSIONS

In MESIA technique, data are collected in a mutual test environment and so it cannot be compared to traditional methods. Thus, MESIA technique is compared with the previous work TANOS technique. Figure 6.5 proves that the existing techniques are not suitable for handling these data. MESIA technique is applied to the Support Vector Machine (SVM) classifier to check the performance. Confusion matrix metrics (precision, recall, accuracy, f1 score) are used to justify the performance of MESIA technique. In figure 6.8 performance metrics chart, X-axis represents the performance metrics and the Y-axis represents the confusion matrix values. Precision is positively predicted values from the collected data. MESIA achieved 98.05% whereas TANOS achieved 97.03%. F1 score is calculated by using precision and recall. MESIA achieved 93.27% F1 score whereas TANOS achieved 92.26%. The recall is the detection possibility of the classifier. MESIA achieved 89.94% recall whereas TANOS achieved 87.14%. Accuracy is the fraction between correct prediction and total prediction. MESIA achieved 89.02% accuracy whereas TANOS achieved 85.18. In MESIA outperforms TANOS in all metrics. TANOS removes sensor errors and it has not handled the features. But TANOS handles the feature selection with the help of seasonally based threshold values. After the MESIA technique, relevant features are selected and irrelevant features are removed from the collected IoT irrigation data. Now the features are reduced from 31 to 5 with the help of mean and seasonal based threshold values. Feature selection is the reason for the rise in classifier accuracy and precision. DaRoN and TANOS handled the noise values and sensor errors. Mean values used for noise removal and neighbour value and parent sensor values are used

to replace the sensor errors. Now, the classifier accuracy and precision are increased a lot while compared with TANOS and DaRoN. Figure 6. 8 shows the performance of TANOS and MESIA in terms of all confusion matrix metrics.

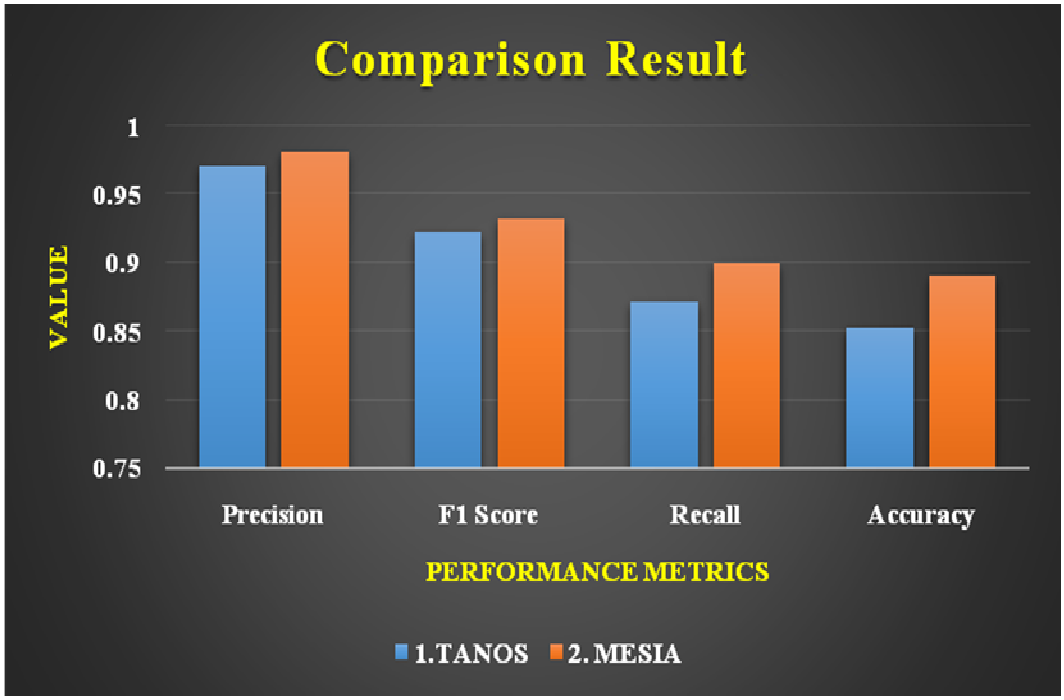


Fig: 6.8: Comparison Result

6.8. FINDINGS AND INTERPRETATIONS

This section explains the experimental results of the TANOS technique and the proposed MESIA technique. It explicates the strength of the MESIA technique and how MESIA technique achieves better results compared to TANOS technique. It also justifies how the MESIA technique is better than the TANOS technique.

The MESIA technique used the statistical measures and seasonal based threshold values for selecting best features in the mutual test IoT environment data. But, TANOS has not focused on feature selection. So, MESIA technique is proposed to handle feature selection.

In this research, the MESIA technique and TANOS technique are applied to the SVM classifier to evaluate the performance of both techniques. MESIA achieved better performance than TANOS in all metrics. So MESIA improved the SVM classifier performance. Table 6.7 shows the comparison of TANOS and MESIA techniques.

Table 6.7: Comparison of TANOS and MESIA

Technique Name	Performance Metrics			
	Precision	**F1 Score**	**Recall**	**Accuracy**
TANOS	0.9703	0.9226	0.8714	0.8518
MESIA	0.9805	0.9327	0.8994	0.8902

6.9. CHAPTER SUMMARY

Traditional feature selection techniques have not focused on continuous features and features with positive and negative correlations. These methods are not capable of handling features collected under a mutual test environment. MESIA technique used the Mean and the Mode values for multivariate filtering and for univariate filtering seasonal-based threshold values and Positive and negative Pearson correlation were used. Combination of both multivariate and univariate filtering yielded good results in terms of improved precision, accuracy, F1 score, recall and classifier accuracy. So, the proposed MESIA technique achieved better performance than the traditional techniques. Thus, the MESIA technique handled continuous features and correlated features and resulted in enhanced classifier accuracy for the collected irrigation dataset.

Chapter – 7

Conclusion

7.1. SUMMARY OF RESEARCH

In proposed Jo's architecture, three preprocessing techniques were proposed, namely

1. DaRoN for noise handling,
2. TANOS for sensor error handling, and
3. MESIA for feature selection.

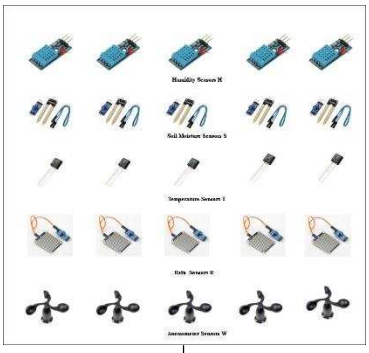
Jo's architecture is discussed in chapter 3. Background knowledge required for the proposed techniques, need for the proposed techniques, methodology of the techniques, and working procedure, were discussed in chapters 4, 5, and 6. In this research work, banana plant irrigation data was collected under IoT mutual test environment. The mutual test environment is an IoT environment that uses the same sensors in a different position to collect data. In this Jo's architecture 5 sensors set was used for data collection, sensors namely, soil moisture sensor, humidity sensor, wind speed (Anemometer) sensor, rain sensor, and temperature sensor. These five sensors set are placed in five different locations in an agriculture field to collect irrigation data. So, there is a total of 25 sensors used for data collection. Sensors name, Number of sensors, units are listed in table 3.3.

Data collection has been enhanced while using a mutual test IoT environment. Traditional IoT application data are not reliable because of the blind spot problem, but this blind spot problem is avoided by the mutual test data collection by using sensors with a wide sensitivity range. Many IoT applications suffer due to missing values and outliers. But outliers are directly eliminated while collecting data under the mutual test environment and missing values are partially eliminated. Due to connection and power problem, missing values exist in the form of sensor errors. The mutual test

environment has two types of sensors used for data collection namely, parent sensor and child sensor. The parent sensor is well placed, has less noise and has a maximum sensitivity range. A child sensor is a supporting sensor to parent sensor that has a minimum sensitivity range and has a huge noise. It covers the area where the parent has a blind spot. There exists a master-slave relationship between parent sensor and child sensor both depend on each other for processing.

Data collection is enhanced in the mutual test environment, but noisy data, namely, repetitive values, null values, sensor errors, etc., are increased due to the number of sensors used. Traditional techniques are not suitable for handling mutual test data. So traditional techniques have to be customized to handle mutual test data. Thus, Jo's architecture proposed three customized techniques to handle mutual test data. Collected data extracted using the Arduino IDE which exported the data into CSV format. This CSV format data is preprocessed using Jo's architecture by using anaconda IDE. Python libraries are used for preprocessing namely Pandas, NumPy, SciPy, Scikit-Learn, and TensorFlow. A machine learning-based SVM classifier is used to check the performance of all three proposed techniques. The working methodology of Jo's architecture is shown in figure 7.1.

Step 1: Data Collection using sensors



Step 2: Data extraction using Arduino

Step 3: Data exported in CSV format

Step 4: Data preprocessing using Anaconda IDE (Python)

Fig 7.1: Methodology of Jo's Architecture

7.2. FEATURES OF THE PROPOSED TECHNIQUES

7.2.1. Jo’s architecture

Jo's architecture is a combination of DaRoN, TANOS, and MESIA techniques. The DaRoN technique is proposed to handle noise data. The TANOS technique is proposed to handle sensor errors that are not focused on by DaRoN. And Finally the MESIA technique is proposed to handle features that are not focused on by TANOS. Now the collected mutual test data is error-free, noise-free and has only relevant features by adding these three techniques. Jo's architecture customized the traditional data mining techniques to handle the collected mutual test data by using DaRoN, TANOS and MESIA techniques. The main advantages of Jo’s architecture are as follows:

- DaRoN technique handles noise in the collected data by using central tendency.
- TANOS technique handles sensor errors by using the neighbor, mean value of child sensors and parent value.
- MESIA technique handles the feature selection by using central tendency and seasonal based threshold values.
- This architecture is suitable for real-time processing.

Table 7.1 shows the details of the proposed techniques in Jo's architecture.

Table 7.1: Details of the proposed techniques

Techniques Name	Problem	Methods/ techniques/ procedures	Novelty
DaRoN	Noise data handling	Central tendency and time stamp value	Combines the three stages of traditional noise removal process into single step Repetitive values are handled
TANOS	Sensor error handling	Central tendency and parent sensor value	Handles sensor errors when both neighbors affected
MESIA	Feature Selection	Central tendency and Seasonal based threshold values	Combine Univariate and Multivariate filtering. Suggest a solution to solve a cloud mask problem

7.2.2. DaRoN Technique

DaRoN technique uses data collected under mutual test IoT environment. Collected irrigation data contain point noise and continuous noise in the form of repetitive values, collisional or inconsistent values and null values. Repetitive values are not a big problem with traditional IoT applications. Because traditional IoT applications deal with fewer sensors while compared with Mutual test IoT applications. But there exist a huge number of repetitive values in the collected data. So, traditional data mining techniques are not capable of handling this many repetitive values. Generally traditional IoT applications suffer due to missing values and outliers. But mutual test environment eliminated outliers and missing values. So traditional data mining technique has to be customized to handle mutual test data. To detect and remove noise in the mutual test IoT irrigation data DaRoN technique is proposed. DaRoN technique combines the three stages of the traditional noise removal process namely, robust, filtering and polishing. Timestamp value used for

robust and filtering. Polishing is done by using a central tendency value. DaRoN has not focused on sensor errors. The main advantages of DaRoN are as follows:

- Traditional noise removal process, namely robust, filtering and polishing combined into a single step by using a timestamp and central tendency value.
- Repetitive values are removed by using time stamp value.
- Inconsistent values are replaced using the central tendency technique.
    - ✓ The parent sensor value used to replace the child sensors noise.
    - ✓ Child sensors, central tendency value which is near to parent value is selected for replacing parent sensor noise.
- Customized the traditional techniques to handle mutual test data.
- DaRoN outperformed all existing techniques in terms of accuracy, precision, F1 score and recall.

7.2.3. TANOS Technique

DaRoN technique ignores sensor errors because sensor errors exist in some rows of the target variable and parent sensor**.** DaRoN has replaced noisy values in the dependent variables by using the central tendency technique. So, the TANOS technique is proposed to handle sensor errors. Collected data affected by the harmless missing values namely, MCAR and MAR. Deletion is the preferable method to handle these errors. But some rows of parent sensors are affected by the sensor errors. So, the imputation technique is selected. TANOS technique customizes the traditional imputation technique to handle mutual test data.Existing sensor error handling techniques sufferers due to both neighbors being affected, matrix value conversion and huge distance between data points. To handle these problems TANOS use neighbor value for replacing instead of matrix conversion. *The parent sensor value* is used to replace

the sensor error value if both *neighbors* are affected. Mean imputation is used when the parent sensor has sensor error. *DaRoN* used mean value so the difference between data points are less so, mean imputation is selected.In this way TANOS techniques handled the sensor errors. The main advantages of TANOS are as follows:

- The matrix conversion problem is avoided by using neighbor sensor values.
- Mean imputation is used to reduce the distance between the data point.
- Parent sensor value is used when both neighbors are affected by sensor errors.
- TANOS outperforms all existing techniques in terms of confusion matrix's metrics namely accuracy, precision, F1 score and recall.

7.2.4. MESIA Technique

Now the data set is error-free and noise-free after applying DaRoN and TANOS techniques. But irrelevant features exist in the collected data set. So, the MESIA technique is proposed to perform the feature selection. There are three types of feature selection methods available namely, filter, wrapper and embedded methods. Among these methods, filter method is selected for feature selection. Because the filter method is classifier independent and easily customizable. Generally, the filter method requires additional supporting techniques to identify variable dependency. So, the MESIA technique uses central tendency and season (Rainy, Strom, Hot, Cold and Cloud mask) based threshold values as supporting techniques. Central tendency is selected because DaRoN and TANOS are already used. Threshold values are defined by the Indian Space Research Centre (ISRC) for each sensor. MESIA Combines two types of filter namely, univariate and multivariate filtering. Initially, the Multivariate filter used central tendency as a supporting technique and perform column-based filtering. Univariate filtering used threshold values as a supporting technique and

perform row and column-based filtering. Multivariate filter reduces the features from 30 to 7 by using the mean and mode values. Univariate filter reduces 7 features into 1 or two by using threshold values. In this way MESIA technique customized the filter method to handle mutual test data. The main advantages of MESIA are as follows:

- Combines multivariate and univariate filtering by using central tendency and threshold values as supporting techniques.
- Performs both column and row-based filtering.
- Among the total 30 features, multivariate filtering selects 7 features and univariate filtering selects either 1 or two but existing methods needs 25 to 30 features.
- MESIA technique makes the IoT application suitable for real-time.

Table 7.2 shows the details of the collected data after applying the proposed techniques.

Table 7.2: Data Details

Technique Name	Data Size
DaRoN	2,98,127
TANOS	2,97,531
MESIA	20,160
*Collected data size is 3,19,520	

7.3. COMPARATIVE ANALYSIS

In Jo's architecture, data are collected in a mutual test environment and so it cannot compared to traditional methods. Because traditional techniques have to be customized to handle mutual test data. Thus the proposed techniques are compared

with the previous work which means TANOS compared with DaRoN, and MESIA compared with TANOS. All techniques are applied to the SVM classifier to check the performance. Confusion matrix metrics (precision, recall, accuracy, f1 score) are used to measure the performance of the proposed techniques. Figure 3.10 showed the performance of the proposed techniques. DaRoN technique achieved 96.05% whereas TANOS and MESIA techniques achieved 97.03% and 98.05% precision i.e., positively predicted values from the collected data. F1 score is calculated using precision and recall measures. F1 score. DaRoN technique achieved 90.27% F1 score whereas TANOS and MESIA techniques achieved 92.26% and 93.27% respectively. The recall is the detection possibility of the classifier. DaRoN technique achieved 85.14 recall value whereas TANOS and MESIA techniques achieved 87.14% and 89.94% respectively. Accuracy is the fraction between correct prediction and total prediction. DaRoN technique achieved 84.18% accuracy whereas TANOS and MESIA techniques achieved 85.18% and 89.02% respectively. The techniques proposed in Jo's architecture increase the machine learning-based SVM classifier performance. Performance metric values are increased from DaRoN to TANOS and TANOS to MESIA. From table 3.7 Jo's architecture achieved 98.05% Precision, 93.27% F1 Score, 89.94% Recall and 0.8902 Accuracy value respectively.

7.4. FINDINGS FROM THE PROPOSED TECHNIQUES

Initially collected data were checked against the recent traditional data mining methods, but the SVM classifier performance is not reaching the satisfiable level. Because repetitive values suffer the traditional methods a lot. So, DaRoN technique was proposed to handle noise in the collected data. DaRoN technique has handled the

noise and increased the performance of the SVM classifier but DaRoN technique has not handled the sensor errors. So, the TANOS technique was proposed to handle sensor errors. TANOS technique has handled the sensor errors and increased the performance of the SVM classifier. TANOS has not focused on feature selection. So the MESIA technique was proposed to handle feature selection and increase the performance of the SVM classifier. Finally by using the proposed three techniques Jo's architecture increased the performance of the SVM classifier to the satisfiable level. The proposed techniques are applicable for any mutual test environment. If the environment same sensors used for data collection in Jo's architecture.If sensors are changed, then DaRoN and TANOS techniques are applicable. But while using the MESIA technique threshold values have to be changed based on sensors.

7.5. LIMITATIONS OF THE PROPOSD TECHNIQUES

Jo's architecture has not focused on time. The only focus is the accuracy of the classifier. Jo's architecture used central tendency, measures to enhance the performance of the SVM classifier. But there are a lot of the latest methods and new versions of classifiers available. If other classifiers and other measures other than the central tendency were used, then it may increase the performance level. The overall accuracy reaches only 89%, which is far better than the existing methods, but it is not the maximum level. If multi-test-based data collection is used, the data collection will be more enhanced. Sensor errors may be avoided if a built-in power supply was used. The proposed techniques are only applicable in any mutual test environment which uses the same sensors used by Jo's architecture.

7.6. RECOMMENDATIONS FOR THE FUTURE WORKS

If the multi-test environment were used then the data collection may be enhanced more, but identification of the parent sensor is difficult in a multi-test environment. The time needed for classification may be focused on in the future. Deep learning-based methods give more accuracy, so new techniques may be proposed based on deep learning. The basic methods, namely central tendency, and Pearson correlation were customized to handle mutual test data collected in Jo's architecture. If new methods were used instead of this, they may increase the accuracy of the classifier. The mutual test environment is less expensive while compared to the multi-test environment. Because, Jo's architecture, data collection costs nearly 1 lakh approximately if the multi-test were used it may cost up to 3 lakhs approximately. But it may remove sensor errors and enhance the data collection. If a built-in power supply may be provided then the sensor error problem may be removed.

References

[**Abd, 21**] Abdulzahra, Suha Abdulhussein, Ali Kadhum M., Al-Qurabat and Ali KadhumIdrees, "Compression-based data reduction technique for IoT sensor networks", Baghdad Science Journal, Vol. 18, No. 1, pp. 0184-0184, 2021.

[**Adi, 20**] Adi, Erwin, Adnan Anwar, Zubair Baig and Sherali Zeadally, "Machine learning and data analytics for the IoT", Neural computing and applications, Vol. 32, No. 20, pp. 16205-16233, 2020.

[**Afs, 21**] Afshan, Nailah and Ranjeet Kumar Rout, "Machine Learning Techniques for IoT Data Analytics", Big Data Analytics for Internet of Things, pp. 89-113, 2021. DOI: https://doi.org/10.1002/9781119740780.ch3.

[Age, 21] Ageed, Zainab Salih, Subhi RM Zeebaree, Mohammed Mohammed Sadeeq, Shakir Fattah Kak, Zryan Najat Rashid, Azar Abid Salih and Wafaa M. Abdullah, "A survey of data mining implementation in smart city applications", Qubahan Academic Journal, Vol. 1, No. 2, pp. 91-99, 2021.

[**Ahm, 22**] Ahmad, MaznahIliyas, YazidSaif, YusriYusof, Md Elias Daud, Kamran Latif and AiniZuhra Abdul Kadir, "A case study: monitoring and inspection based on IoT for milling process", The International Journal of Advanced Manufacturing Technology, Vol. 118, No. 3, pp. 1305-1315, 2022.

[**Ait, 19**] Ait Issad H., Aoudjit R., and Joel JPC Rodrigues, "A comprehensive review of Data Mining techniques in smart agriculture", Engineering in Agriculture, Environment and Food, Vol. 12, No. 4, pp. 511-525, 2019.

[**Alc, 19**] Alcalde Barros A., García Gil D., García S. and Herrera F., "DPASF: a flink library for streaming data pre-processing", Big Data Analytics, Vol.4, No. 1, pp. 1-17, 2019.

[**Alq, 19**] Al Qurabat A. K. M., Abou Jaoude C. and Idrees A. K., "Two tier data reduction technique for reducing data transmission in IoT sensors", International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 168-173, 2019. DOI: doi: 10.1109/ IWCMC.2019. 8766590.

[**Ana, 21**] Anachkova, Maja, Simona Domazetovska, Zlatko Petreski and Viktor Gavriloski, "Design of low-cost wireless noise monitoring sensor unit based on IoT concept", Journal of Vibro engineering, Vol. 23, No. 4, pp. 1056-1064, 2021.

[**And, 20**] Andersen D. L., Ashbrook C. S. A. and Karlborg N. B., "Significance of big data analytics and the internet of things (IoT) aspects in industrial development, governance and sustainability", International Journal of Intelligent Networks, Vol. 1, pp. 107-111, 2020. DOI: https://doi.org/10.1016/j.ijin.2020.12.003.

[**Arm, 17**] Armina R., Mohd Zain A., Ali N. A. and Sallehuddin R., "A Review On Missing Value Estimation Using Imputation Algorithm", Journal of Physics: Conference Series, Vol. 892, pp. 1-12, 2017. doi:10.1088/ 1742-6596/892/1/012004.

[Ash, 09] Ashton Kevin, "That 'internet of things' thing", RFID journal, Vol. 22, No. 7, pp. 97-114, 2009.

[**Asi, 18**] Asiya Sulthana and Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOT Enabled Telecardiology System", International Journal of Engineering & Technology, Vol. 7, No. 2, pp. 74-78, 2018.

[**Ass, 17**] Assahli S., Berrada M. and Chenouni D., "Data pre-processing from the Internet of Things: Comparative study", Wireless Technologies, Embedded and Intelligent Systems (WITS), pp. 1-4, 2017. DOI: 10.1109/ WITS.2017. 7934676.

[**Aur, 19**] Aurora Gonzalez-Vidal, Fernando Jimenez, Antonio F. and Gomez-Skarmeta, "A methodology for energy multivariate time series forecasting in smart buildings based on feature selection", Energy and Buildings, Vol. 196, pp. 71-82, 2019, DOI: https://doi.org/10.1016/ j.enbuild.2019.05.021.

[**Ayv, 21**] Ayvaz, Serkan and KorayAlpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time", Expert Systems with Applications, Vol. 173, pp. 1-10, 2021. DOI: https://doi.org/10.1016/j.eswa.2021.114598.

[**Bev, 03**] Bevington Philip R. and Keith Robinson D., "Data reduction and error analysis", McGrawa Hill, New York, pp. 1-12, 2003.

[**Bev, 93**] Bevington P.R., Robinson D.K., Blair J.M., Mallinckrodt A.J. and McKay S., "Data reduction and error analysis for the physical sciences" Computers in Physics, Vol.7, No. 4, pp. 415-416, 1993.

[**Bor, 14**] Borgia, Eleonora, "The Internet of Things vision: Key features, applications and open issues", Computer Communications, Vol. 54, pp.1-31, 2014. DOI: https://doi.org/10.1016/j.comcom.2014.09.008.

[**Cal, 17**] Calmon F. P., Wei D., Vinzamuri B., Ramamurthy K. N. and Varshney K. R., "Optimized pre-processing for discrimination prevention", International Conference on Neural Information Processing Systems, pp. 1-10, 2017.

[**Can, 18**] Cano J.C., Berrios V., Garcia B. and Toh C.K., "Evolution of IoT: an industry perspective", IEEE Internet of Things Magazine, Vol. 1, No. 2, pp. 12-17, 2018.

[**Cha, 19**] Chao Xu, Howard H. Yang, Xijun Wang and Tony Q. S. Quek., "On Peak Age of Information in Data Preprocessing enabled IoT Networks", IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6, 2019. DOI:10.1109/wcnc.2019.8885690, 2019.

[**Cha, 21**] Chan R. K. C., Lim J. M.Y., Parthiban R., "A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system", Expert Systems with Applications, Vol. 171, pp. 1-14, 2021. DOI: https://doi.org/10.1016/j.eswa. 2021. 114573.

[**Cho, 01**] Choh Man Teng, "A Comparison of Noise Handling Techniques", FLAIRS Conference, pp. 269-273, 2001.

[**Cim, 20**] Cimmino, Andrea, Maria Poveda-Villalon and Raul Garcia-Castro, "ewot: A semantic interoperability approach for heterogeneous iot ecosystems based on the web of things", Sensors, Vol. 20, No. 3, pp. 1-19, 2020.

[**Cui, 18**] Cui L., Yang S., Chen F., Ming Z., Lu N. and Qin J., "A survey on application of machine learning for Internet of Things", Vol. 9, No. 8, pp. 1399-1417, 2018.

[**Cur, 19**] Curley C., Krause R. M., Feiock R. and Hawkins C. V., "Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database", Urban Affairs Review, Vol. 55, No. 2, pp. 591-615, 2019.

[**Dab, 19**] Dabbakuti JR., Jacob A., Veeravalli VR. and Kallakunta RK., "Implementation of IoT analytics ionospheric forecasting system based on machine learning and ThingSpeak", Sonar, & Navigation, Vol. 14, No. 2, pp. 341-347, 2019.

[**Dac, 19**] Dachyar M., Teuku Yuri M. Zagloel and Ranjaliba Saragih L., "Knowledge growth and development: internet of things (IoT) research, 2006–2018", Heliyon, Vol. 5, No. 8, pp. 1-14, 2019.

[**Dev, 20**] Devi Lakshmi R. and Kalaivani V., "Machine learning and IoT-based cardiac arrhythmia diagnosis using statistical and dynamic features of ECG", The Journal of Supercomputing, Vol. 76, No. 9, pp. 6533-6544, 2020.

[**Dey, 21**] Dey Indrakshi and Shama Siddiqui, "Wavelet Transform for Signal Processing in Internet-of-Things (IoT)", Wavelet Theory, London: IntechOpen, pp. 183-202, 2021. DOI: 10.5772/intechopen.95384.

[**Du, 20**] Du, Qingbo, Faming Yin and Zongchen Li, "Base station traffic prediction using XGBoost-LSTM with feature enhancement", IET Networks, Vol. 9, No. 1, pp. 29-37, 2020.

[**Ege, 17**] Egea, Santiago, Albert Rego Manez, Belen Carro, Antonio Sanchez-Esguevillas and Jaime Lloret, "Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments", IEEE Internet of Things Journal, Vol. 5, No. 3, pp. 1616-1624, 2017, 2017.

[Elb, 21] Elbadawi, Moe, Simon Gaisford and Abdul W. Basit, "Advanced machine-learning techniques in drug discovery", Drug Discovery Today, Vol. 26, No. 3, pp. 769-777, 2021.

[**Evg, 18**] Evgeniy Latyshev, "Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance", Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL), Vol. 2277, pp. 226-231, 2018,

[**Far, 20a**] Farahani, Bahar, Mojtaba Barzegari, Fereidoon Shams Aliee and Khaja Ahmad Shaik, "Towards collaborative intelligent IoT eHealth: From device to fog, and cloud", Microprocessors and Microsystems, Vol. 72, pp.1-16, 2020. DOI: https://doi.org/10.1016/j.micpro.2019.102938.

[**Far, 20b**] Farooq, Muhammad Shoaib, Shamyla Riaz, Adnan Abid, Tariq Umer and Yousaf Bin Zikria, "Role of IoT technology in agriculture: A systematic literature review", Electronics, Vol. 9, No. 2, pp. 1-41, 2020.

[**Fer, 14**] Fernandez Perez M. P. and Gonzalez Navarro F. F., "Non-deterministic Local Search Methods for Feature Selection: An Experimental Study", International Conference on Artificial Intelligence, pp. 69-74, 2014. DOI:10.1109/micai.2014.16.

[**Gar, 16**] Garcia Luis P.F. Carvalho Andre C.P.L.F. de and Lorena Ana C., "Noise detection in the meta-learning level", Neurocomputing, Vol. 176, pp. 14-25, 2016. DOI:10.1016/j.neucom.2014.12.100.

[**Gar, 18**] Garcia Gil D., Luengo J., Garcia S. and Herrera F., "Enabling Smart Data: Noise filtering in Big Data classification", Information Sciences, Vol. 479, pp.135-152, 2018. DOI:10.1016/j.ins.2018.12.002.

[**Gee, 21**] Geeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm. (Accessed on 19 December, 2021)

[**Gil, 16**] Gil Press, 2016, The evolving IT landscape. https://www.student. unsw. edu.au/how-do-i-cite-electronic-sources. (Accessed on 12 November, 2021)

[**Gio, 20**] Giouroukis, Dimitrios, Alexander Dadiani, Jonas Traub, Steffen Zeuch and Volker Markl, "A survey of adaptive sampling and filtering algorithms for the internet of things", ACM International Conference on Distributed and Event-based Systems, pp. 27-38. 2020. DOI: https://doi.org/10.1145/3401025.3403777.

[**Gla, 20**] Glaroudis, Dimitrios, Athanasios Iossifides and Periklis Chatzimisios, "Survey, comparison and research challenges of IoT application protocols for smart farming", Computer Networks, Vol. 168, pp. 1-36, 2020. DOI: https://doi.org/10.1016/j.comnet.2019.107037.

[**Gon, 19**] Gonzalez Vidal, Aurora, Fernando Jimenez and Antonio F. Gomez Skarmeta, "A methodology for energy multivariate time series forecasting in smart buildings based on feature selection", Energy and Buildings, Vol. 196, pp. 71-82, 2019. DOI: https://doi.org/ 10.1016/ j.enbuild. 2019.05.021.

[**Goo, 16**] GoodfellowI., Bengio Y. and Courville A. J. D. l.,"Machine learning basics", Vol. 1, No. 7, pp. 98-164, 2016.

[**Gop, 18**] Gopika N. and Meena kowshalaya M.E. A., "Correlation Based Feature Selection Algorithm for Machine Learning", International Conference on Communication and Electronics Systems (ICCES), pp. 692-695, 2018. DOI:10.1109/cesys.2018.8723980.

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

[**Gui, 21**] Guillen-Navarro, Miguel A., Raquel Martínez-Espana, Belen Lopez and Jose M. Cecilia, "A high-performance IoT solution to reduce frost damages in stone fruits", Concurrency and Computation: Practice and Experience, Vol. 33, No. 2, pp. 1-14, 2021.

[**Gup, 19**] Gupta S. and Gupta A., "Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review", Procedia Computer Science, Vol. 161, pp. 466–474, 2019. DOI:10.1016/j.procs. 2019.11.146.

[**Had, 20**] Hadeed, Steven J., Mary Kay O'Rourke, Jefferey L. Burgess, Robin B. Harris and Robert A. Canales, "Imputation methods for addressing missing data in short-term monitoring of air pollutants", Science of The Total Environment, Vol. 730, pp. 1-7, 2020. DOI: https://doi.org/ 10.1016/j.scitotenv.2020.139140.

[**Hir, 15**] Hira Z. M. and Gillies D. F, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", Advances in Bioinformatics, pp.1-13, 2015. DOI: http://dx.doi.org/10.1155/ 2015/198363.

[**Hos, 20**] Hossain T., Ahad M., Rahman A. and Inoue S, "A method for sensor-based activity recognition in missing data scenario", Sensors, Vol. 20, No. 14, pp. 1-23, 2020.

[**Hui, 20**] Huifeng, Wang, Seifedine Nimer Kadry and Ebin Deni Raj, "Continuous health monitoring of sportsperson using IoT devices based wearable technology", Computer Communications, Vol.160, pp.588-595, 2020. DOI: https://doi.org/ 10.1016/j.comcom.2020. 04.025.

[**Huq, 18**] Huque M. H., Carlin J. B., Simpson J. A. and Lee K. J, "A comparison of multiple imputation methods for missing data in longitudinal studies", BMC medical research methodology, Vol. 18, No.1, pp. 1-16, 2018.

[**IBM, 21**] IBM. https://www.ibm.com/in-en/cloud/learn/machine-learning. (Accessed on 19 December, 2021).

[**Jam, 19**] Jamshed Huma, Sadiq Ali Khan, Muhammad Khurram, Syed Inayatullah and Sameen Athar, "Data Preprocessing: A preliminary step for web data mining", 3c Tecnologia: glosas de innovacion aplicadas a la pyme, Vol. 8, No. 1, pp.206-221, 2019.

[**Jan, 21a**] Jane V. A. and Arockiam L., "Survey on IoT Data Preprocessing", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol. 12, No. 9, pp. 238-244, 2021.

[**Jan, 21b**] Jane V. A. and Arockiam L., "DaRoN: A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency", Annals of the Romanian Society for Cell Biology, Vol. 25, No.2, pp. 3197-3203, 2021.

[Kot, 18] Kotha, Harika Devi and Mnssvkr Gupta V., "IoT application: a survey", Int. J. Eng. Technology, Vol.7, No. 2.7, pp. 891-896, 2018.

[**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.

[**Kum, 19**] Kumar S., Tiwari P. and Zymbler M., "Internet of Things is a revolutionary approach for future technology enhancement: a review", Journal of Big Data, Vol. 6, No.1, pp. 1-21, 2019.

[**Kum, 20**] Kumar, Adarsh, Krishnamurthi Rajalakshmi, Saurabh Jain, Anand Nayyar and Mohamed Abouhawwash, "A novel heuristic simulation-optimization method for critical infrastructure in smart transportation systems", International Journal of Communication Systems, Vol. 33, No. 11, pp. 1-34 , 2020.

[**Kun, 21**] Kun Tan, Sun Sanmin, Du Liangzong and Zhou Shaoliang, "Design of an Intelligent Irrigation System for a Jujube Orchard based on IoT", INMATEH-Agricultural Engineering, Vol. 63, No. 1, pp. 189-198, 2021.

[**Kwa, 17**] Kwak S. K. and Kim J. H., "Statistical data preparation: management of missing values and outliers", Korean Journal of Anesthesiology, Vol. 70, No. 4, pp.1-5, 2017.

[**Lan, 17**] Lan K., Fong S., Song W., Vasilakos A. and Millham R, "Self-Adaptive Pre-Processing Methodology for Big Data Stream Mining in the Internet of Things Environmental Sensor Monitoring", Symmetry, Vol. 9, No. 10, pp. 1-17, 2017.

[**Lee, 19**] Lee Minseok, Jihoon An and Younghee Lee, "Missing-Value Imputation of Continuous Missing Based on Deep Imputation Network Using Correlations among Multiple IoT Data Streams in a Smart Space", IEICE Transactions on Information and Systems, Vol. 102, No. 2, pp. 289-298, 2019.

[**Len, 02**] Lenzerini M., "Data integration: A theoretical perspective", ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 233-246, 2002. DOI: https://doi.org/ 10.1145/ 543613. 543644.

[**Li, 20**] Li You, Yuan Zhuang, Xin Hu, Zhouzheng Gao, Jia Hu, Long Chen and Zhe He, "Toward location-enabled IoT (LE-IoT): IoT positioning techniques, error sources, and error mitigation", IEEE Internet of Things Journal, Vol. 8, No. 6, pp. 4035-4062, 2020.

[**Lin, 19**] Lin, Yi-Bing, Yun-Wei Lin, Jiun-Yi Lin and Hui-Nien Hung, "Sensor Talk: An IoT device failure detection and calibration mechanism for smart farming", Sensors, Vol. 19, No. 21, pp. 1-19, 2019.

[**Lin, 20**] Li Linchao, Bowen Du, Yonggang Wang, Lingqiao Qin and Huachun Tan, "Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model", Knowledge-Based Systems, Vol.194, pp. 1-25, 2020. DOI:10.1016/j.knosys.2020.105592.

[**Liu, 20a**] Liu, Yuehua, Tharam Dillon, Wenjin Yu, Wenny Rahayu and Fahed Mostafa. "Noise removal in the presence of significant anomalies for Industrial IoT sensor data in manufacturing", IEEE Internet of Things Journal, Vol. 7, No. 8, pp. 7084-7096, 2020.

[**Liu, 20b**] Liu, Yuehua, Tharam Dillon, Wenjin Yu, Wenny Rahayu and Fahed Mostafa, "Missing value imputation for Industrial IoT sensor data with large gaps", IEEE Internet of Things Journal, Vol. 7, No. 8, pp. 6855-6867, 2020.

[**Liu, 21**] Liu, Li-Wei, Mohd Hasmadi Ismail, Yu-Min Wang and Wen-Shin Lin, "Internet of Things based Smart Irrigation Control System for Paddy Rice Field", AGRIVITA, Journal of Agricultural Science, Vol. 43, No. 2, pp. 1-10, 2021.

[**Luf, 21**] Luftensteiner, Sabrina, Michael Mayr and Georgios Chasparis, "Filter-Based Feature Selection Methods for Industrial Sensor Data: A Review", Big Data Analytics and Knowledge Discovery, Springer, pp. 242-249, 2021. DOI: https://doi.org/10.1007/978-3-030-86534-4_23.

[**Mao, 19**] Mao Yifei and Yuansheng Yang, "A Wrapper Feature Subset Selection Method Based on Randomized Search and Multilayer Structure", BioMed Research International, pp. 1-9, 2019. DOI: https://doi.org/ 10.1155/2019/9864213.

[**Med, 20**] Medapati Prema Kumar, Tejo Murthy P. H. S. and Sridhar K. P., "LAMSTAR: For IoT-based face recognition system to manage the safety factor in smart cities", Transactions on Emerging Telecommunications Technologies, Vol. 31, No. 12, pp. 1-15, 2020.

[**Mia, 16**] Miao Jianyu and Lingfeng Niu, "A Survey on Feature Selection", Information Technology and Quantitative Management, Vol. 91, pp. 919-926, 2016. DOI: https://doi.org/10.1016/j.procs.2016.07.111.

[**Moh, 19**] Mohtashami, Mohammad and Mahdi Eftekhari, "A hybrid filter-based feature selection method via hesitant fuzzy and rough sets concepts", Iranian Journal of Fuzzy Systems, Vol. 16, No. 2, pp. 165-182, 2019.

[**Mor, 19**] Morais C. M. de., Sadok D. and Kelner J., "An IoT sensor and scenario survey for data researchers", Journal of the Brazilian Computer Society, Vol. 25, No. 1, pp. 1-17, 2019.

[**Nat, 16**] Natarajasivan D. and Govindarajan M., "Filter Based Sensor Fusion for Activity Recognition using Smartphone", International Journal of Computer Science and Telecommunications, Vol. 7, No.5, pp. 26-31, 2016.

[**Niz, 20**] Nizetic, Sandro, Petar Solic, Diego Lopez-de-Ipina Gonzalez-de and Luigi Patrono, "Internet of Things (IoT): Opportunities, issues and challenges towards a smart and sustainable future", Journal of Cleaner Production, Vol. 274, pp.1-77, 2020. DOI: https://dx.doi.org/ 10.1016/ j.jclepro.2020.122877.

[**Nou, 19**] Noura, Mahda, Mohammed Atiquzzaman and Martin Gaedke, "Interoperability in internet of things: Taxonomies and open challenges", Mobile Networks and Applications, Vol. 24, No. 3, pp. 796-809, 2019.

[**Pap, 18**] Papageorgiou Grigorios, Stuart W. Grant, Johanna JM Takkenberg and Mostafa M. Mokhles, "Statistical primer: how to deal with missing data in scientific research?", Interactive CardioVascular and Thoracic Surgery, Vol. 27, No. 2, pp.153–158, 2018.

[**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, 19**] Pathak Abhijit, Mohammad AmazUddin, Md Jainal Abedin, Karl Andersson, Rashed Mustafa and Mohammad Shahadat Hossain, "IoT based smart system to support agricultural parameters: A case study", Procedia Computer Science, Vol. 155, pp. 648-653, 2019. DOI: https://doi.org/10.1016/j.procs.2019.08.092.

[**Pen, 19**] Peng Tao, Sana Sellami and Omar Boucelma, "IoT data imputation with incremental multiple linear regression", Open Journal of Internet Of Things (OJIOT), Vol. 5, No. 1, pp. 69-79, 2019.

[**Pet, 17**] Peter Wlodarczak, Mustafa Ally and Jeffrey Soar, "Data Mining in IoT", Knowledge Management of Web Social Media (KMWSM '17), pp. 1100-1103, 2017. DOI: https://doi.org/10.1145/3106426.3115866.

[**Poi, 21**] Point J., 2021. https://www.javatpoint.com/data-integration-in-data-mining. Accessed on 19 December, 2021.

[**Rad, 21**] Radhakrishnan G., Srinivasan K., Maheswaran S., Mohanasundaram K., Palanikkumar D. and Vidyarthi A., "A deep-RNN and meta-heuristic feature selection approach for IoT malware detection", Materials Today: Proceedings, pp. 2214-7853, 2021. DOI: https://doi.org/10.1016/j.matpr.2021.01.207.

[Ram, 17] Ramirez Gallego S., Krawczyk B., Garcia S., Wozniak M. and Herrera F., "A survey on data preprocessing for data stream mining: Current status and future directions", Neuro computing, Vol. 239, pp.39-57, 2017. DOI: https://doi.org/10.1016/j.neucom.2017.01.078.

[**Ran, 21**] Rani, Pooja, Rajneesh Kumar and Anurag Jain, "Multistage model for accurate prediction of missing values using imputation methods in

heart disease dataset", Innovative data communication technologies and application, Springer, pp. 637-653, 2021. DOI: https://doi.org/ 10.1007/978-981-15-9651-3_53.

[**Rod, 21**] Rodriguez, Jhonn Pablo, Ana Isabel Montoya-Munoz, Carlos Rodriguez-Pabon, Javier Hoyos and Juan Carlos Corrales, "IoT-Agro: A smart farming system to Colombian coffee farms", Computers and Electronics in Agriculture, Vol. 190, pp. 106442, 2021. DOI: https://doi.org/10.1016/j.compag.2021.

[**Rua, 19**] Ruan, Junhu, Yuxuan Wang, Felix Tung Sun Chan, Xiangpei Hu, Minjuan Zhao, Fangwei Zhu, Baofeng Shi, Yan Shi and Fan Lin, "A life cycle framework of green IoT-based agriculture and its finance, operation, and management issues", IEEE communications magazine, Vol. 57, No. 3, pp. 90-96, 2019.

[**Sae, 17**] Saez, Jose A., Mikel Galar, Julian Luengo and Francisco Herrera, "INFFC: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control", Information Fusion, Vol. 27, pp. 19-32, 2016. DOI: https://doi.org/10.1016/j.inffus.2015.04.002.

[**San, 18**] Sanyal Sunny and Zhang Puning, "Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications", IEEE Access, Vol. 6, pp. 67830–67840, 2018. DOI: https://doi.org/ 10.1109/ ACCESS.2018.2878640.

[**Sav, 18**] Savaliya, Akshat, Aakash Bhatia and Jitendra Bhatia, "Application of Data Mining Techniques in IoT: A Short Review", Int. J. Sci. Res. Sci. Eng. Technol , Vol.4, No. 2, pp. 218-223, 2018.

[**Sha, 16**] Shahzadi, Raheela, Muhammad Tausif, Javed Ferzund and Muhammad Asif Suryani, "Internet of things based expert system for smart agriculture", Int. J. Adv. Comput. Sci. Application, Vol. 7, No. 9, pp. 341-350, 2016.

[**Sha, 17**] Shanthamallu, Uday Shankar, Andreas Spanias, Cihan Tepedelenlioglu and Mike Stanley, "A brief survey of machine learning methods and their sensor and IoT applications", International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1-8, 2017. DOI: https://doi.org/ 10.1109/IISA.2017.8316459.

[**Sha, 20**] Shah, Devarshi, Jin Wang and Q. Peter He, "Feature engineering in big data analytics for IoT-enabled smart manufacturing–comparison between deep learning and statistical learning", Computers & Chemical Engineering, Vol. 141, pp. 1-22, 2020. DOI: https://doi.org/ 10.1016/ j.compchemeng.2020.106970.

[**Sin, 22**] Sinha, Bam Bahadur and R. Dhanalakshmi, "Recent advancements and challenges of Internet of Things in smart agriculture: A survey", Future Generation Computer Systems, Vol. 126, pp. 169-184, 2022. DOI: https://doi.org/10.1016/j.future.2021.08.006.

[**Sto, 18**] Stonebraker, Michael and Ihab F. Ilyas, "Data Integration: The Current Status and the Way Forward", IEEE Data Eng. Bull., Vol. 41, No. 2, pp. 3-9, 2018.

[**Su, 19**] Su Shen, Yanbin Sun, Xiangsong Gao, Jing Qiu and Zhihong Tian, "A correlation-change based feature selection method for IoT equipment anomaly detection", Applied sciences, Vol. 9, No. 3, pp. 1-14, 2019.

[**Sud, 18**] Sudha Rani K. and Nageshwar Rao D., "A comparative study of various noise removal techniques using filters" Research & Reviews: Journal of Engineering and Technology, Vol. 7, No. 2, pp. 47-52, 2018.

[**Sul, 18**] Sulthana, Asiya and Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOT enabled telecardiology system", Int. J. Eng. Technology, Vol. 7, No. 2, pp. 74-78, 2018.

[**Sun, 18**] Sun Guanglu, Jiabin Li, Jian Dai, Zhichao Song and Fei Lang, "Feature selection for IoT based on maximal information coefficient", Future Generation Computer Systems, Vol. 89, pp. 606-616, 2018.

[**Sun, 20a**] Sundararajan, Karthik and Anandhakumar Palanisamy, "Multi-rule-based ensemble feature selection model for sarcasm type detection in twitter", Computational intelligence and neuroscience, pp. 1-17, 2020. DOI: https://doi.org/10.1155/2020/2860479.

[**Sun, 20b**] Sunhare P., Chowdhary R. R. and Chattopadhyay M. K., "Internet of Things and Data Mining: An Applications Oriented Survey", Journal of King Saud University - Computer and Information Sciences, pp. 1-26, 2020. DOI: https://doi.org/10.1016/j.jksuci.2020.07.002.

[**Swa, 16**] Swati Jain, Kalpana Jain and Naveen Chodhary, "A Survey Paper On Missing Data In Data Mining", International Journal Of Innovations In Engineering Research And Technology [Ijiert], Vol. 3, No. 12, pp. 45-50, 2016.

[**Swe, 17**] Swetha G. and Ramya G., "Survey paper on Big data imputation and Privacy algorithms", International Research Journal of Engineering and Technology (IRJET), Vol.04 , No. 07, pp. 3441-3443, 2017.

[**Tad, 19**] Tadist, Khawla, Said Najah, Nikola S. Nikolov, Fatiha Mrabti and Azeddine Zahi, "Feature selection methods and genomic big data: a systematic review", Journal of Big Data, Vol. 6, No. 1, pp. 1-24, 2019.

[**Tal, 17**] Talavera Jesus Martin, Luis Eduardo Tobon, Jairo Alejandro Gomez, Maria Alejandra Culman, Juan Manuel Aranda, Diana Teresa Parra, Luis Alfredo Quiroz, Adolfo Hoyos and Luis Ernesto Garreta, "Review of IoT applications in agro-industrial and environmental fields", Computers and Electronics in Agriculture, Vol. 142, pp. 283-297, 2017. DOI: https://doi.org/10.1016/j.compag.2017.09.015.

[**Tan, 17**] Tang F. and Ishwaran H., "Random forest missing data algorithms Statistical Analysis and Data Mining", The ASA Data Science Journal,Vol. 10, No. 6, pp. 363–377, 2017.

[**Tao, 19**] Tao Peng, Sana Sellami and Omar Boucelma, "IoT Data Imputation with Incremental Multiple Linear Regression", Open Journal of Internet of Things (OJIOT), Vol. 5, No. 1, pp. 1-2, 2019.

[**Teh, 20**] Teh, Hui Yie, Kempa-Liehr, Andreas W, and Wang Kevin I-Kai, "Sensor data quality: a systematic review", Journal of Big Data, Vol. 7, No. 1, pp. 11-60, 2020.

[**Ter, 21**] Terms, B. https:/ /binaryterms.com /data -reduction. (Accessed on 19 December, 2021).

[**Thi, 21**] Thiam, Fatoumata, Maissa Mbaye and Alexander M. Wyglinski, "Generic Reliability Analysis Model of IoT: Agriculture use case", Vehicular Technology Conference (VTC2021-Spring), IEEE , pp. 1-5, 2021. DOI: https://doi.org/10.1109/VTC2021-Spring 51267. 2021. 9448791.

[**Tka, 21**] Tkachenko, Roman, Ivan Izonin, Ivanna Dronyuk, Mykola Logoyda and Pavlo Tkachenko, “Recovery of missing sensor data with grnn-based cascade scheme”, International Journal of Sensors Wireless Communications and Control, Vol. 11, No. 5, pp. 531-541, 2021.

[**Tod, 86**] Tody Doug, “The IRAF data reduction and analysis system." In Instrumentation in astronomy VI, Vol. 627, pp. 733-748, 1986. DOI: https://doi.org/10.1117/12.968154.

[**Tsa, 19**] Tsamardinos I., Borboudakis G., Katsogridakis P., Pratikakis P. and Christophides V. A., “greedy feature selection algorithm for Big Data of high dimensionality”, Machine Learning, Vol. 108, No. 2, pp.149-202, 2019.

[**upG, 21**] upGrad.https://www.upgrad.com/blog/methods-of-data-transformation- in-data-mining/ (Accessed on 19 December, 2021).

[**Vad, 21**] Vadivelu R., Santhakumar G., Boopana V. P., Dharani S., Dhivya Bharathi S. and Balasubramaniam D., “Improved Air Quality Testing Using Kalman Filter Based on IoT Data Fusion for Smart Cities” Journal of Physics: Conference Series, Vol. 1916, No. 1, pp. 1-7, 2021.

[**Van, 21**] Vandana Charaliparambu Pathayapuram and Ajeet Annarao Chikkamannur, “An Ameliorated Ensemble Approach for IoT Resource Feature Selection Based on Discriminating and Service Relevance Criteria”, International Journal of Intelligent Engineering and Systems, Vol.14, No.3, pp. 435-446, 2021.

[**Ver, 20**] Veras M. B., Mesquita D. P., Mattos C. L. and Gomes J. P, “A sparse linear regression model for incomplete datasets”, Pattern Analysis and Applications, Vol. 23, pp. 1-11, 2020. DOI: https://doi.org/ 10.1007/ s10044-019-00859-3.

[**Vil, 20**] Villa Henriksen, Andres, Gareth TC Edwards, Liisa A. Pesonen, Ole Green and Claus Aage Gron Sorensen, “Internet of Things in arable farming: Implementation, applications, challenges and potential”, Biosystems Engineering, Vol. 191, pp. 60-84, 2020. DOI: https://doi.org/10.1016/j.biosystemseng.2019.12.013.

[**Wah, 18**] Wah, Yap Bee, Nurain Ibrahim, Hamzah Abdul Hamid, Shuzlina Abdul-Rahman and Simon Fong, “Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy”, Pertanika Journal of Science & Technology, Vol. 26, No. 1, pp. 329-340, 2018.

[**Wan, 20**] Wang, Jianzhou, Ying Wang, Zhiwu Li, Hongmin Li and Hufang Yang, “A combined framework based on data preprocessing, neural networks and multi- tracker optimizer for wind speed prediction”, Sustainable Energy Technologies and Assessments, Vol. 40, pp. 1-19, 2020. DOI: https://doi.org/10.1016/j.seta.2020.100757.

[**Wat, 21**] Watanabe Futa, “Wireless Sensor Network Localization Using AoA Measurements with Two-Step Error Variance-Weighted Least Squares”, IEEE Access, Vol. 9, pp. 10820-10828, 2021.

[**Wei, 18**] Weiss, Matthew, Michael S. Wiederoder, Randy C. Paffenroth, Eric C. Nallon, Collin J. Bright, Vincent P. Schnee, Shannon McGraw, Michael Polcha and Joshua R. Uzarski, “Applications of the Kalman filter to chemical sensors for downstream machine learning”, IEEE Sensors Journal, Vol.18, No. 13, pp. 5455-5463, 2018.

[**Wij, 20**] Wijesekara WM and Liyanage L., "Comparison of imputation methods for missing values in air pollution data: a case study on Sydney Air Quality index", Advances in intelligent systems and computing, Vol.1130, pp. 257-269, 2020. DOI: https://doi.org/10.1007/978-3-030-39442-4_20.

[**Yen, 19**] Yen Neil Y., Jia-Wei Chang, Jia-Yi Liao, and You-Ming Yong, "Analysis of interpolation algorithms for the missing values in IoT time series: a case of air quality in Taiwan", The Journal of Supercomputing, Vol.76, No. 8, pp. 6475-6500, 2019.

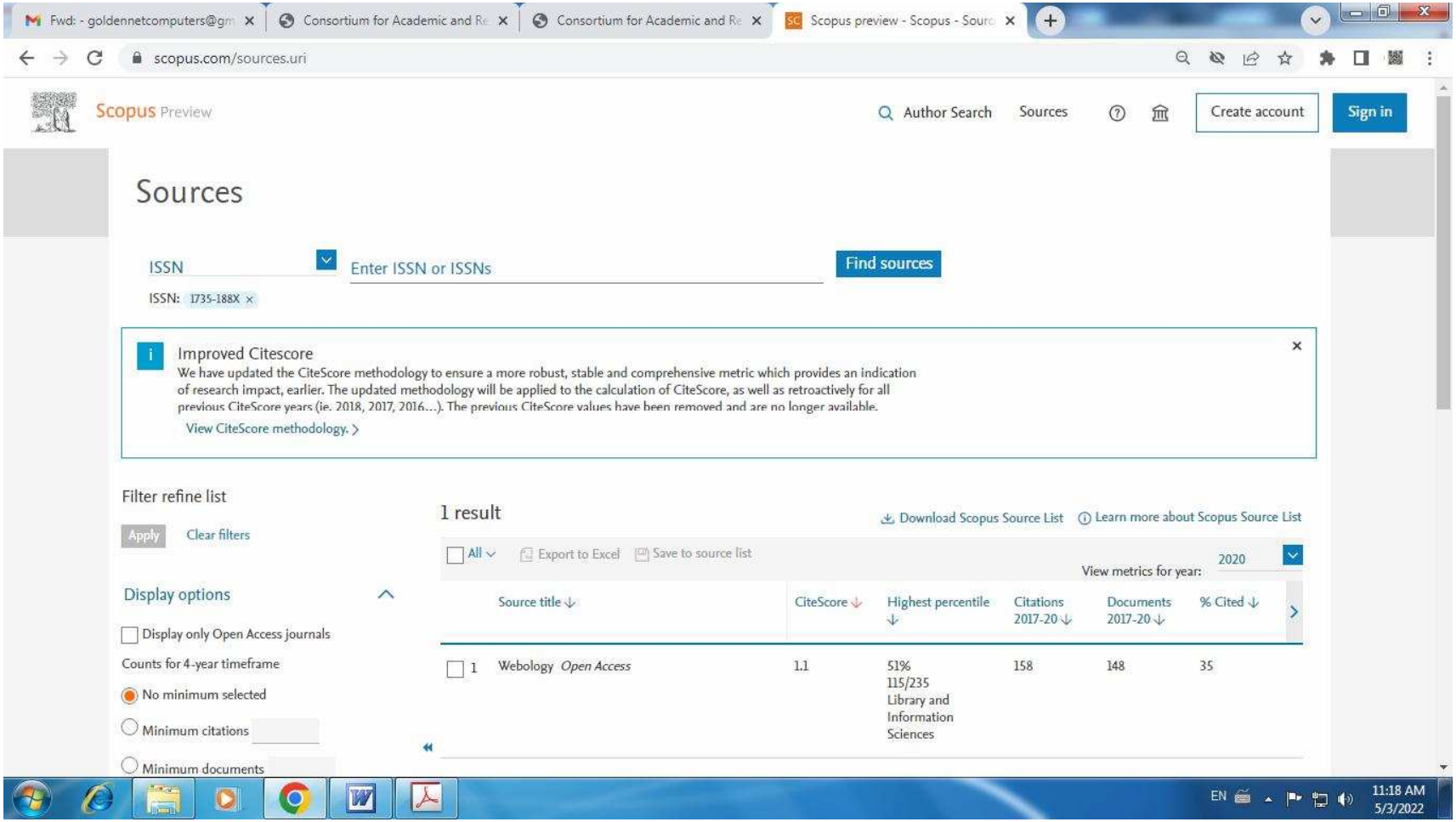
[**Yi, 19**] Yi Bing Lin, Yun Wei Lin, Jiun Yi Lin and Hui Nien Hung, "SensorTalk: An IoT Device Failure Detection and Calibration Mechanism for Smart Farming", Sensors, Vol. 19, No. 21, pp. 1-19, 2019.

[**Yu, 21**] Yu, Wenjin, Yuehua Liu, Tharam Dillon, WennyRahayu and Fahed Mostafa, "An Integrated Framework for Health State Monitoring in a Smart Factory Employing IoT and Big Data Techniques", IEEE Internet of Things Journal, Vol. 3, No. 3, pp. 2443-2454, 2022.

[**Zen, 15**] Zena M. Hira and Duncan F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", Advances in Bioinformatics, pp. 1-13, 2015. DOI: https://doi.org/10.1155/2015/198363.

[**Zho, 19**] Zhong Y., Fong S., Hu S., Wong R. and Lin W., "A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", Sensors, Vol. 19, No. 20, pp.1-16, 2019.

Photocopies of Papers Published in the International Journals


Fwd: - goldennetcomputers@gm
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Scopus Preview
Author Search
Sources
Create account
Sign in
Sources
ISSN
Enter ISSN or ISSNs
Find sources
ISSN: 1735-188X
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
Webology Open Access
1.1
51% 115/235 Library and Information Sciences
158
148
35
EN
11:18 AM
5/3/2022

Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Iot Data Preprocessing - A Survey

V.A. Jane[1] , Dr. L. Arockiam[2]

[1, 2]Department of computer Science, St. Joseph's College (Affiliated to Bharathidasan University**)**, Trichy, Tamilnadu, India 62001.

Abstract
The Internet of Things (IoT) is a rapidly evolving system in engineering and science. The sensors utilized in numerous sectors output a significant quantity of data. As a result, several consumers have a clear desire for efficient knowledge from these vast databases. This enormous dataset is far from flawless; it has several flaws (like distortion, inconsistent data, and anomalies) and is unsuitable for investigation due to the risk of inaccurate results. As a result, data preprocessing is a necessary approach for these information. Data preprocessing is a crucial and necessary phase, with the primary purpose of using procedures to filter, refine, repair, and enhance the raw information. The purpose of this study is to conduct a survey of IoT data preprocessing and methodologies. This article analyses current data preprocessing studies in the IoT environment, as well as the history of IoT data preprocessing and review articles of sophisticated data preprocessing approaches. The image clearly depicts the categorization of different preprocessing methodologies and procedures. Preprocessing cleaning, conversion, minimization, and integration methods are discussed. Furthermore, strategies for implementing such ideas in IoT data preprocessing are presented. IoT approaches for data preparation in diverse applications are listed. Lastly, difficulties and obstacles are explored that will be valuable in future research.

Key Words IoT, Preprocessing, Data Cleaning, Noise handling

Introduction

The Internet of Things (IoT) is a system of items that are linked to the Internet. It's a powerful automated and intelligence platform with applications in a variety of sectors and distinctive adaptability and capabilities in every provided setting (for example, agriculture and medicine) [1]. Being linked to the Internet, one may gather information and transfer it over the web, acquire data from the web, or do both. In the IoT, sensors and other devices produce data tremendously. Such information are transported to the cloud for processing, evaluation or modeling and to construct software applications. Big data analysis is a very essential method

of finding insights from such information [2]. Data preparation is essential before evaluating the data since it contains various flaws like missing data, distortion, and inconsistency. One of the most important aspects of the knowledge discovery process is data preparation [3]. Low-quality information may weaken the efficacy of subsequent learning algorithms. As a result, limiting the effect on reliability improves the dependence of following automated findings and improves judgments via the use of appropriate processing techniques. Data transformation, data reduction, data standardization, data cleansing, and data integration are some of the strategies used [5]. By separating complicated continual feature sets and choosing and deleting undesired and noisy characteristics, such strategies minimize the information. Throughout this procedure, the actual construction of the data must be preserved while a more acceptable size is achieved. Quick training of learning approaches, sophisticated generalization abilities, and improved comprehension and convenient analysis of the results are among the advantages of data processing [6]. The purpose of this study is to conduct a survey of data preprocessing, its approaches, and current data preprocessing achievements. The following is the framework of this work: Firstly, data pretreatment ideas in IoT contexts (part 2), as well as data preprocessing methodologies. Section 3 explains the methods used in numerous IoT-based applications, and Section 5 brings this project to a close.

Related work

Physical sensor faults that arise throughout the data gathering process were examined by Hui et al., [7]. This article describes wide range of physical sensor discrepancies, error - detecting processes, and error - correcting methods, as well as the variations between them. Principal component analysis (PCA) and Artificial Neural Network (ANN) were the best error-detection and rectification procedures.

Mathew et al., [8] examined Kalman filter, z-scoring, and moving Average filter processing approaches. To begin, the chemical sensor data was cleaned using pre-processing steps. Following that, the information is cleaned and assessed utilizing classification techniques like Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), and Support Vector Classifier (SVC). Lastly, the different preprocessing approaches' efficiencies were evaluated. Among them, the Kalman filter approach was shown to produce superior results than the rest.

For a higher-dimensional Microarray tumor sample, Zena et al. [9] analyzed feature-selection and feature-extraction approaches. In the micro array set of data, the researcher addressed the implications of redundant and irrelevant attributes. The significance of dimension reduction, as well as its benefits and downsides, also were explored.

The method of data transfer in an IoT context was outlined by Chao et al.,[10]. Although a pretreatment approach was employed to reduce transmission time and increase processing speed, this research centered on the latter.

Evgeniy [11] suggested a processing system for sensor information. Several processing strategies that are appropriate for the proposed framework have been discovered. This framework used streaming sensor data from the Univariate time series dataset.

Filter-based monitoring solution for IoT environment was suggested by Natarajasivan et al., [12]. Accelerometers, position sensors, vision sensors, audio sensors, temperature sensors, and directional sensor readings were all used in this project. The obtained information from such sensing devices was processed utilizing Kalman filter, and the performance was evaluated employing SVM. The suggested scheme took longer to complete.

Cleber et al., [13] conducted a review of all IoT application journal articles since 2015. The authors assigned a value to the IoT application depending on how it was used. When contrasted to other apps, smart home applications are commonly employed by investigators. In addition, the sensor used in intelligent devices is explored.

Rajalakshmi et al., [14], addressed the concept of IoT in intelligent systems and summarized sensor data-collection issues like data aggregation, extensibility, data fusion, de-noising, variability, data anomaly analysis, real-time computation, and missing data imputation. The authors discuss the IoT data analytics procedure using a drone for a traffic-monitoring scheme and described how cloud, fog, and edge computing are used in IoT to enhance the analytics platform.

Data gathering, cleansing, data aggregation, data migration to the cloud, and data processing were all discussed by David et al., [15] in their examination of data management issues in the IoT context. AI, machine learning, deep learning, and data mining are some of the enhanced data-processing innovations explored by the researcher.

Karinaer al., [16] gave an overview on processing strategies as well as data mining-related challenges. The essential ideas of data mining, as well as processing approaches and challenges, were thoroughly addressed. It also explored recommendations for future and provided alternative ideas.

Data preprocessing methodologies for the big data era were suggested by Garca et al., [17]. The critical elements of data processing were discussed, as well as the existing key problems. In addition, various data preprocessing techniques for text mining were examined, including discretization and normalization, extraction of features, feature selection, feature indexers and encoders, and other methods. The importance of large data preparation was also stressed.

In the area of data mining, Jayaram et al., [18] published a survey on data preparation strategies. The major goal was to find answers to different data preparation issues. The authors concentrated on data cleaning techniques such as filtering, imputation, hybrid approach, wrapper techniques, and embedded methodologies. Every technique's procedure and applications were detailed with instances. Distortion and data management, particularly, were discussed, as well as instructions on how to identify and handle it. Lastly, the difficulties encountered while performing data cleansing in various domains were shown.

Several large data processing strategies were explored by Huma Jamshed et al., [19] to cleanse input for subsequent mining and analytical jobs. The basic phases of data preparation, such as data filtering, data conversion, data reduction, and data aggregation, were first described. Following that, architecture for internet data preparation was presented, with every step described in detail. Lastly, the model was applied to the basic textual data, and processing stages such as removal of noise, tokenization, and normalization were completed.

Taxonomy of preprocessing techniques

Data processing is the process of preparing actual data to be used in data mining [20]. Actual data is noisier, has incomplete data, it includes a lot of uncertain information, and it's enormous. Such factors influence the quality of the data to deteriorate during the mining or modeling process, resulting in poor results. As a result, information should be improved before it can be mined or modeled. This is referred to as data preprocessing. There are several approaches for doing such operations in order to make the data appropriate for analysis. IoT data preprocessing techniques are shown in figure1.

Data Cleaning

Data cleaning [21] can be defined as the process of eliminating the erroneous and missing part in the data. The process of handling these noisy and missing values can be achieved by various ways.

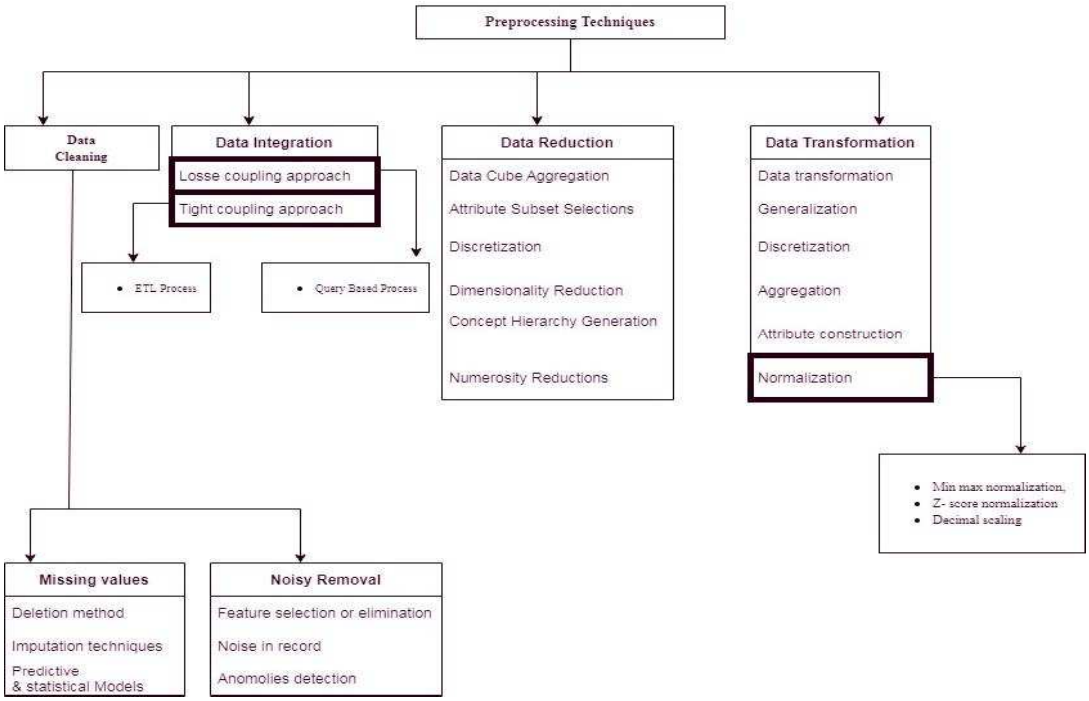


Fig 1: Taxonomy of IoT data preprocessing

Deal with missing values

Now a days, Missing values are present in all types of datasets that from existing, industrial and devices [22]. They have different reasons like manual data entry procedures, equipment Errors and misalignments. Such datasets require a pre-processing level to prepare and clean a data for effective and sufficient knowledge extraction process. There are machine learning algorithms and packages that can automatically detect and manipulate missing data, but it is recommended to manually replace missing data with analysis and coding techniques [23]. First, the types of missing have to understand and basically missing is classified into three types that are missing at random (MCAR), missing at Random (MAR) and missing not at Random (MNAR) [24]. To deal all these types of missing values using various methods like removing data [25], imputation techniques [26], multiple imputation techniques [27], statistical or predictive models [28]. In the removing method, missing values handle through list wise, pairwise and dropping variables techniques. The goal of any imputation technique is to create a complete dataset, which can be used for machine learning. Some of the imputation techniques are, deductive imputation, mean/ median /mode imputation, random sampling imputation, regression imputation, multiple imputation [29]. Predictive and statistics model also used to impute the missing data. Most commonly used in this process are, linear regression, Random forest, k nearest neighbor, expectation maximization and sensitivity analysis [30].

Deletion / Removing Method

Deletion / removing methods used to remove the missing values using various approaches that are following,

List-wise deletion

The most common approach to handle missing data, it simply omit the data that with missing values after analyze the remaining data. This process also called as complete case analysis [31]. If the samples are large and the assumption of MCAR is satisfied, list wise deletion is in reasonable strategy. However, when a large sample is not available, or the assumption of MCAR is not satisfied, the list wise approach is not the optimal strategy. This approach establishes the bias if it does is not fulfill the MCAR.

Pairwise Deletion

Pairwise deletion technique remove missing observations only and the existing variables are analyzed [32]. If there is no data elsewhere in the data set, existing values are used. This approach uses all of the observed information, so it saves more information than list wise deletion technique. Pairwise deletion also called as Available Case Analysis (ACA). Pairwise deletion is known to be less dependent on MCAR or MAR data. However, if there are many

missing observations, the analysis is flawed. The problem with pairwise elimination is that even if one takes the available cases, one cannot compare the analyzes because each time the model is different.

Dropping Variables

Dropping variables approach removes a variable or column from a dataset if it contains more missing values [33]. This approach is performs depend on the situation there is no rule to handle this approach and requires a proper analysis of the data before the variable is dropped all together. This should be the last option to test whether the model improves performance after the variable is removed.

Dealing with Noisy data

Noisy data is an unwanted data item, feature, or record that does not help explain the feature, or the relationship between the feature and the target [34]. Noisy data can affect the algorithms to find the patterns in the data. Noise can be classified into three types [35]. Noise 1 is anomalies in some data items, noise 2 is features that don't help to the target like irrelevant or weak features, and noise 3 is which records that do not follow the form or relationship that like the rest of the records. If the noise is in the features, feature selection or elimination techniques to best for handling noise in the features this includes filter method, wrapped method and embedded methods. For handling noise in records, k fold validation and manual methods are used in basic. Unsupervised methods are used to detect anomalies in data items. Some of them are, density based anomaly detection, cluster based anomaly detection and SVM based anomaly detection [36].

Data integration

Data integration **[37-38]** is one important techniques in preprocessing which combines data from different source and giving users an integrated view of this data. Data sources may contain databases, data cubes, or flat files. One of the most popular implementations of data integration is the creation of a company's data warehouse. Mainly, Data integration is done through two main approaches known as the "tight coupling approach" and "loose coupling approach"[39]. Tight coupling defines Data from various sources that combined into one place by the process of Extraction, Transformation and Loading. Single physical location provides a balanced interface for querying data and ETL process provide identical data warehouse. Loose coupling data exists only in real source databases. In loose coupling, virtual mediation schema takes an interface from the user to the query, converts and sends it to a source database for getting result. As well as there are many "adapters" or "wrappers" in the mediation schema that can be reconnected to the source systems and bring the data to the front end.

Data reduction

Over the past decades, data generation and storage in data bases or data warehouses has increased. So, these amounts data can take a very long time to perform data analysis and mining process. Data reduction **[**40-41**]** techniques can be used to obtain a data set, which are very small in size but yield the same analytical results. Traditional, data reduction approaches [42-43] are Data cube aggregation, Attribute subset selections, Dimensionality reduction, Discretization and concept hierarchy generation and Numerosity reductions. Data cube aggregation used to construct a data in simple form. It applied on the data and form a data cubes. Attribute subset selection technique remove irrelevant, weakly features or redundant features. This process can be achieved by various statistical and computational methods like filter methods, wrapper methods, and embedded methods. Dimensionality reduction is the reduction technique to reduce the size of dataset. This process used to reduce the number of random variables to be considered by obtaining the set of the principal variable. Dimensionality reduction reduces the amount of data by eliminating outdated or unwanted features using techniques that includes PCA, backward feature elimination, forward feature construction, and discriminant methods. Since real data is replaced with real data, with mathematical models or a small representation of data like parameters or non-parametric method such as clustering, sampling and histogram. Discretization & Concept Hierarchy Operation techniques are used to change the raw data values for the attributes by a range or by more conceptual conditions. This is a form of numerical reduction, which is very useful for the automatic generation of concept sequences. Discretization techniques follows two ways namely top down discretization and bottom up discretization. Concept hierarchies for numeric data that includes techniques are binning, histogram analysis and clustering.

Data transformation

Data transformation [44] is the process of converts' data from one format to another format. Data transformation includes smoothing, aggregation, discretization, attribute construction, normalization and generalization. Smoothing used to remove noise from a dataset though various algorithms and highlight the significant features in a dataset. Data normalization involves converting all data variables into a specific range. Techniques used for normalization min max normalization, z- score normalization and decimal scaling**.**

Conclusion

Big data is currently widely used in a variety of fields, including academia, agribusiness, medicine, organizations, and web mining. Studying from such vast amounts of data is both an exciting and difficult undertaking. Information acquired from massive quantities of data offers tremendous prospects and has the ability to alter several industries. However, since big data contains imperfections such as noise and incomplete information, it may reduce the effectiveness and reliability of decision-making. As a result, data refining is required. In the case of larger data settings, this work presents a systematic flow of research on data

preparation strategies. The principles of data preparation were discussed, as well as literature evaluations pertaining to data preprocessing approaches. The image clearly demonstrated the categorization of different pre-processing methodologies and procedures. Preprocessing, cleansing, transformations, reductions, and collaboration with methodologies and procedures were all shown. On a variety of applications, data preparation methods were tabulated. Finally, difficulties and concerns that should be addressed in the future were discussed.

References

[1] Bramer, Max. “Data for data mining”, In Principles of data mining”, pp. 9-19. Springer, London, 2016.

[2] Alasadi, Suad A., and Wesam S. Bhaya.”Review of data preprocessing techniques indata mining, Journal of Engineering and Applied Sciences 12, no. 16 (2017): 4102-4107, 2017.

[3] Cordón, Ignacio, Julián Luengo, Salvador García, Francisco Herrera, and Francisco Charte. & quot; Smart data: Data preprocessing to achieve smart data in r.& quot; Neuro computing 360, 1-13, 2019.

[4] Hu, Hanqing, and Mehmed Kantardzic. & quot; Smart preprocessing improves data streammining. & quot; In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp.1749-1757. IEEE, 2016.

[5] Shi, F.; Li, Q.; Zhu, T.; Ning, H., “A survey of data semantization in internet of things”, Sensors, 18, 313, 2018.

[6][24] Shah, S. H., &Yaqoob, I, “A survey: Internet of Things (IOT) technologies, applications and challenges“, IEEE Smart Energy Grid Engineering SEGE). doi:10.1109/sege.2016.7589556, 2016.

[7] Teh, Hui Yie, Kempa-Liehr, Andreas W, Wang, Kevin I-Kai, "Sensor data quality: a systematic review", Journal of Big Data, 7(1), 11–60, 2020, doi: 10.1186/s40537-020-0285-1

[8] Weiss, Matthew, Wiederoder, Michael S, Paffenroth, Randy C, Nallon, Eric C, Bright, Collin J, Schnee, Vincent P, McGraw, Shannon; Polcha, Michael, Uzarski, Joshua R, "Applications of the Kalman Filter to Chemical Sensors for Downstream Machine Learning", IEEE Sensors Journal, (), 1–1, 2018, doi:10.1109/JSEN.2018.2836183

[9] Hira, Z. M., &Gillies, D. F, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", Advances in Bioinformatics, 1–13, 2015,doi:10.1155/2015/198363

[10] Xu, C., Yang, H. H., Wang, X., &Quek, T. Q. S, "On Peak Age of Information in Data Preprocessing enabled IoT Networks", IEEE Wireless Communications and Networking Conference (WCNC), 2019, doi:10.1109/wcnc.2019.8885690

[11] Evgeniy Latyshev, "Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance", Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL'2018), 226-231, 2018.

[12] D. Natarajasivan and M. Govindarajan, "Filter Based Sensor Fusion for Activity Recognition using Smartphone", International Journal of Computer Science and Telecommunications Volume 7, Issue 5, 2016.

[13] Morais, C. M. de, Sadok, D., & Kelner, J, "An IoT sensor and scenario survey for data researchers", Journal of the Brazilian Computer Society, 25(1), doi:10.1186/s13173-019-0085-7, 2019.

[14] Rajalakshmi Krishnamurthi, Adarsh Kumar, Dhanalekshmi Gopinathan, Anand Nayyar, and Basit Qureshi, "An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques", Sensors, 20, 6076; doi:10.3390/s20216076.

[15] Gil, D., Johnsson, M., Mora, H., & Szymanski, J, "Advances in Architectures, Big Data, and Machine Learning Techniques for Complex Internet of Things Systems", Complexity, 1–3, doi:10.1155/2019/4184708,2019.

[16] Gibert, Karina, Miquel Sànchez–Marrè, and Joaquín Izquierdo. "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining." AI Communications 29, no. 6 (2016): 627-663.

[17] García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. "Big data preprocessing: methods and prospects." Big Data Analytics 1, no. 1 (2016): 9.

[18] Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1-5. IEEE, 2017.

[19] Jamshed, Huma& Khan, M. &Khurram, Muhammad & Inayatullah, Syed &Athar, Sameen. (2019). Data Preprocessing: A preliminary step for web data mining. 206-221. 10.17993/3ctecno.2019.specialissue2.206-221.

[20] [21] Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey, "International Conference on Intelligent Sustainable Systems (ICISS)", ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.

[21] Syaifudin, Yan Watequlis, and Dwi Puspitasari. "Twitter data mining for sentiment analysis on people's feedback against government public policy." MATTER: International Journal of Science and Technology 3, no. 1, 2017.

[22] Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic" Journal of Intelligent Learning Systems and Applications, 9, no. 01, 1, 2017.

[23] Yadav, Madan Lal, and Basav Roychoudhury. "Handling missing values: A study of popular imputation packages in R", Knowledge-Based Systems, 160, 104-118, 2018.

[24] Gomes, Harold, "Evaluation of Patterns of Missing Prices in CPI Data.", 2018.

[25] Huang, Min-Wei, Wei-Chao Lin, Chih-Wen Chen, Shih-Wen Ke, Chih-Fong Tsai, and William Eberle. "Data preprocessing issues for incomplete medical datasets" Expert Systems, 33, no. 5, 432-438, 2016.
[26] Wang, Guang C., Kenny C. Gross, and Dieter Gawlick. "Missing value imputation technique to facilitate prognostic analysis of time-series sensor data." U.S. Patent Application 16/005,495, filed December 12, 2019.
[27] Chhabra, Geeta, Vasudha Vashisht, and Jayanthi Ranjan, "A comparison of multiple imputation methods for data with missing values", Indian Journal of Science and Technology, 10, no. 19 (2017): 1-7.
[28] Alexandropoulos, Stamatios-Aggelos N., Sotiris B. Kotsiantis, and Michael N. Vrahatis. "Data preprocessing in predictive data mining." The Knowledge Engineering Review, 34, 2019.
[29] Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)" Artificial Intelligence Review, 53, no. 2 (2020): 1487-1509, 2020.
[30] Gad, Ibrahim, and B. R. Manjunatha. "Performance evaluation of predictive models for missing data imputation in weather data." In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1327-1334. IEEE, 2017.
[31] Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1-5. IEEE, 2017.
[32] Al-Utaibi, Khaled A., and El-Sayed M. El-Alfy. "Intrusion detection taxonomy and data preprocessing mechanisms", Journal of Intelligent & Fuzzy Systems 34, no. 3 (2018): 1369-1383, 2018.
[33] Zulkepli, Fatin Shahirah, Roliana Ibrahim, and Faisal Saeed. "Data preprocessing techniques for research performance analysis." In Recent Developments in Intelligent Computing, Communication and Devices, pp. 157-162. Springer, Singapore, 2017.
[34] Misra, Puneet, and Arun Singh Yadav, "Impact of Preprocessing Methods on Healthcare Predictions" Available at SSRN 3349586 (2019).
[35] Nayak, Arjun Srinivas, A. P. Kanive, Naveen Chandavekar, and R. Balasubramani. "Survey on pre-processing techniques for text mining." International Journal Of Engineering And Computer Science, ISSN (2016): 2319-7242, 2016.
[36] Kumar, HM Keerthi, and B. S. Harish. "Classification of short text using various preprocessing techniques: An empirical evaluation" In Recent Findings in Intelligent Computing Techniques, pp. 19-30. Springer, Singapore, 2018.
[37] Hui, Jingya, Lingli Li, and Zhaogong Zhang. "Integration of big data: a survey." In International Conference of Pioneering Computer Scientists, Engineers and Educators, pp. 101-121. Springer, Singapore, 2018.

[38] Samoilova, Evgenia, Florian Keusch, and Tobias Wolbring. "Learning analytics and survey data integration in workload research." Zeitschrift für Hochschulent wicklung. Special Edition: Learning Analytics: Implications for Higher Education 12, no. 1 (2017): 65-78.

[39] Osial, Phillip, Kalle Kauranen, and Emdad Ahmed. "Smartphone recommendation system using web data integration techniques." In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-5. IEEE, 2017.

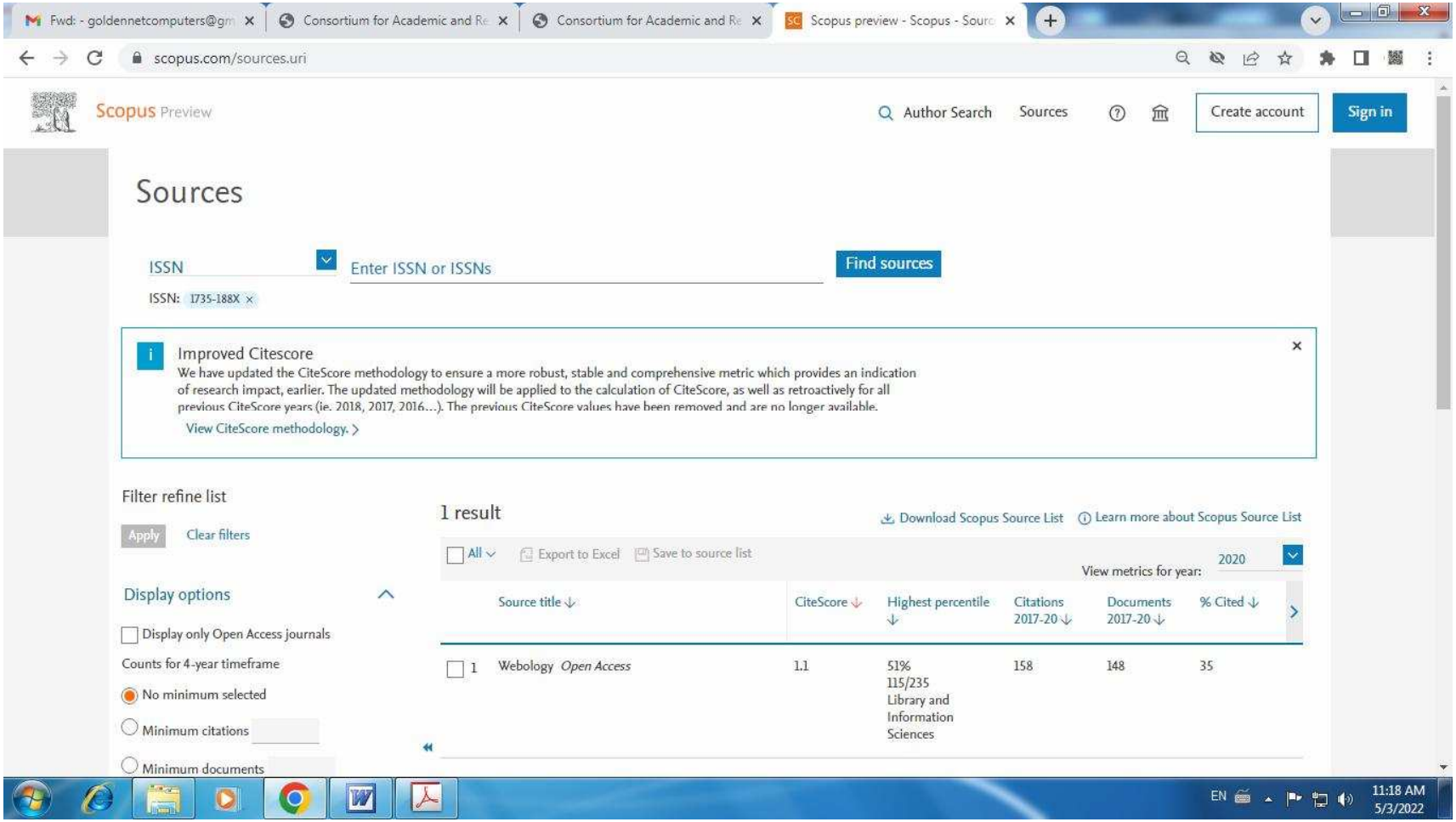
[40] ur Rehman, Muhammad Habib, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U. Khan. "Big data reduction methods: a survey." Data Science and Engineering 1, no. 4 (2016): 265-284, 2016.

[41] ur Rehman, Muhammad Habib, Victor Chang, Aisha Batool, and Teh Ying Wah. "Big data reduction framework for value creation in sustainable enterprises." International Journal of Information Management 36, no. 6 (2016): 917-928, 2016.

[42] Weng, Jiaying, and Derek S. Young, "Some dimension reduction strategies for the analysis of survey data", Journal of Big Data, 4, no. 1 (2017): 1-19, 2017.

[43] Jindal, Priyanka, and Dharmender Kumar. "A review on dimensionality reduction techniques", International journal of computer applications 173, no. 2 (2017): 42-46, 2017.

[44] Jarmin, Ron S., and Amy B. O'Hara. "Big data and the transformation of public policy analysis", Journal of Policy Analysis and Management, 35, no. 3 (2016): 715-721, 2016.


Fwd: - goldennetcomputers@gm
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Scopus Preview
Author Search
Sources
Create account
Sign in
Sources
ISSN
Enter ISSN or ISSNs
Find sources
ISSN: 1735-188X
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
Webology Open Access
1.1
51% 115/235 Library and Information Sciences
158
148
35
EN
11:18 AM
5/3/2022

Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

CID: Central Tendency Based Noise Removal Technique For Iot Data

V. A. Jane[1*] , Dr. L. Arockiam[2],

[1*, 2] Department of Computer Science, St. Joseph's College (Affiliated to Bharathidasan University), Tiruchirappalli.

Abstract
The Internet of Things (IoT) is a crucial technique that enables well-organized and reliable solutions for various areas' development. Agriculture is one of the most worried IoT areas, with IoT-based solutions being used to automate the management and evaluation system with the least amount of human participation. Every second, a large-scale IoT-based agricultural ecosystem creates a tremendous quantity of data. The agro-production ecosystem is complicated, and there are several inconsistencies in the raw data that assessment and mining cannot be directly tracked. This research presents a strategy called Detection and Removal of Noise (CID) to deal with these anomalies in IoT agriculture data. Utilizing measurements of central tendency, the suggested approach eliminates null values, incorrect values, repeating values, unfinished values, and inappropriate values. In addition, a comparison of current noise reduction strategies was carried out, and the effectiveness was assessed using the Support Vector Machine (SVM) classification. To improve accuracy of classification, noisy data is removed in this suggested investigation. The CID approach will help improve the quality of data obtained in agricultural settings.

Key Words Noise, Data cleaning, IoT, Pre processing, Noise removal, Smart Agriculture.

Section I: Introduction

IoT is a popular technology that uses its capabilities to make numerous applications smarter [1]. Collecting information in the agricultural area was previously a challenging operation, particularly in surveillance devices, but the Internet of Things (IoT) eliminates all of those demanding parts with the use of sensors. Sensors play a critical role in data collecting and create massive amounts of data on a daily basis. There are missing values, distortion, anomalies, and duplicated values in this information [2]. If any of the aforementioned are contained in the obtained data, the output quality may suffer. Noise is one of the most significant, and it is described as useless data such as corrupted values, repeating values, incorrect values, null values, and so on. These issues arise as a

result of IoT-related issues like connection errors, detection errors, and collisions [3]. Several issues may arise throughout the analysis procedure if the dataset includes noisy data.

Point noise and continuous noise are two different forms of noise. The Point noise deviates sharply from the rest of the data. As a result, this might be clearly spotted. Since the divergence increases from point to point, continuous noise is hard to detect. The mean, median, and mode approaches are employed to remove this sort of noise. The incidences in the data may also be used to characterize noises. Class noise is defined as noise that happens in the class column. When noise appears in the attribute column, it is referred to as attribute noise. Unlike class noise, attribute noise is more destructive since it influences the data directly. As a result, noise in the database may influence the analytics model's accuracy [4]. As a result, data pre-processing is required.

Data cleansing, data integration, data conversion, and data reduction are some of the sorts of pre-processing procedures [5]. This article is about noise reduction, which is part of the data cleaning procedure. Section II examines similar works in the relevant domain, Section III defines the procedure of the proposed study, Section IV summarizes the findings and discussion, and Section V ends the work.

Section II: Related Works

The function of data mining in IoT was discussed by Peter et al.,[6]. This paper covered all of the techniques, methodologies, and processes connected to data mining in relation to different IoT systems. It also explained the significance of data management in smart settings.

To manage data in the Data Stream Mining (DSM), Kun et al. [7] suggested a clustering-based particle swarm optimization (CPSO) technique. Data segmentation was done using the sliding window approach, and variable partition was done using Statistical Feature Extraction (SFX). The suggested method was tested with five different kinds of IoT data sets (Home, Gas, Ocean, and Electricity). The results were analyzed, and the suggested method enhanced efficiency while increasing computational overhead and the overfitting issue.

Various pre-processing strategies for mining and analytical activities were explored by Huma Jamshed et al., [8]. The essential techniques of data pre-processing, such as cleaning the data, data conversion, data reduction, and data aggregation, were detailed in this study. The researcher offered a method for doing so, which he demonstrated using a simple text data case study. Noise reduction, tokenization, and normalization were all addressed by the suggested method. According to the findings, modern approaches such as machine learning boosted the efficacy of pre-processing.

In an IoT-enabled Telecardiology platform, Asiya et al. [9] analyzed the accuracy of noise cancelling approaches. LMS (Least Mean Square), NLMS (Normalized Least Mean Square), CLLMS (Circular Leaky Least Mean Square), and VSS-CLLMS (Variable Step Size CLLMS) were the approaches used for comparison. Removal of baseline wander (BW) (minimum frequencies in ECG (Electro Cardio Gram)). The VSS-CLLMS approach produced a high SNRI (Signal to Noise Ratio Improvement). The researchers concentrated on filtering strategies for ECG data pre-processing.

Using an outlier detection methodology, Liu et al. [10] suggested a method for dealing with distortion in IoT data. Using a sliding window and analytical measures, the suggested methodology calculated the variation and divergence. It also recognized distortion in the dataset based on neighbourhood activity, making the task of removing incorrect data more challenging if an issue was found in the continuous neighbourhood. The identifying procedure took longer in this case.

Wang et al., [11] established a framework for pre-processing and forecasting wind data. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) approach was used to eliminate noise from wind data, and MTO (multi-tracker optimizer) was employed to find errors in this suggested study. Eventually, neural network layers were used to construct the model. The suggested CEEMDAN approach was only acceptable for small datasets, and when bigger datasets were investigated, the mean error worsened.

Sanyall et al., [12] suggested a technique to deal with IoT sensor data authenticity issues (noise, null values, anomalies, and repetition). The suggested technique was divided into two sections: the first dealt with data aggregation using the clustering technique, and the other with data pre-processing employing resilient dominant subspace calculation and monitoring techniques. Outliers rose as a result of the randomized outputs created by dominating subspace selections, lowering overall effectiveness.

Sáez et al., [13] proposed the Iterative Class Noise Filter (INFFC), which iteratively merged several classifiers for distortion detection. The filtering approach was developed to define the distortion by removing the noise detection step from every iteration.

Garcia et al., [14] used an aggregation of noise filtering approaches to enhance noise identification. Meta Learner (MTL) was developed as a method for reducing duplicate data and eliminating unnecessary data in a database. Meta features were produced from damaged dataset for this purpose, and a meta-learning framework that forecasted noisy information was developed as a result.

Section III: CID Proposed Methodology

Irrigation systems in agriculture needs regular monitoring without human interference. The suggested CID approach collects information from IoT devices and saves it in the cloud to automate this procedure. The obtained data is then pre-processed utilizing central tendency metrics, and the efficiency of the pre-processed data is evaluated using a Support Vector Machine (SVM) classification. Robust (identification of every analytical flaws to make the data standardized), Filtering (utilizing different approaches to reduce noise), and Polishing (changing incorrect values) are the three steps of classic noise treatment techniques [15]. The suggested CID is unique in that it blends the three stages into one to provide a noise - free data.

Pre-defined criteria are used for robust and filtration, and measurements of central tendency are used for refining.

The pre-processed agricultural data generates a noise - free cleansed data as a result of the suggested work's methodology that improves the classifier's effectiveness.

Phase I: Sensing Layer

The first stage focuses on data collecting in an agricultural setting utilizing different IoT devices. The humidity sensor, temperature sensor, soil moisture sensor, wind speed sensor, and rain sensor are among the 5 sensors employed. Sensors are installed in various locations and are linked to the cloud. Every sensor monitors the surroundings in its own way and captures information in real time. A humidity sensor captures information about the amount of water in the air. This information would be valuable in assessing whether or not watering is required. The proportion of water in the soil is measured using a soil moisture sensor. Both humidity and soil moisture sensor data are used in this study to determine irrigation recommendations. Temperature sensors are often used to detect temperature levels on a regular basis. The rain sensor is used to measure the amount of rain falling. This sensor's principal function is to turn off the whole irrigation system during intense rainfall. The data collected from the sensors is periodically gathered and transferred to the cloud for more analysis.

Phase 2: Storing Layer

The storing layer is the second layer, and it is used to store information. Information may be kept locally, but cloud storage is the most efficient way to manage enormous amounts of information. As a result, the information is stored in the cloud using the suggested method. Several open-source clouds are accessible; one of them is the Think Speak cloud server that offers an open-source computation paradigm for storing and retrieving data over the network. A service provider is responsible for maintaining, operating, and managing information. In Think Speak, an account is formed and developed with numerous fields for storing information like soils, moisture, heat, and rainfall. The information is then sent to the pre-processing phase.

Phase 3. Pre processing Layer

The proposed noise reduction approach is applied in this layer. The metrics of central tendency are used in this unique method. Conventional noise reduction methods consist of three stages: robust, filtering, and polishing. The suggested CID approach, on the other hand, integrates all three stages into a single stage by employing measurements of central tendency, resulting in improved effectiveness. To substitute repetitious and null entries, the suggested approach uses the nearest mean values. To eliminate Point Noise, the Nearest Mode value is used. All changes are completed in accordance with the time details (Td). The measurements of central tendency are mentioned below.

Mean (μ) $= \frac{\text{sum of all elements}}{\text{Total number of Elements}}$

Median (M) $= L + h\frac{((fm - f1))}{((fm - f1) - (fm - f2))}$

Mode (Z) $= \frac{(n+1)}{2}$

Let L = {L_1,L_2 .. L_n}, where, L_1, L_2 …L_n are different locations.

Every location contains numerous sensors that are T_n, S_n, H_n, R_n and W_n in which n indicates count of locations, T_n – Temperature sensor, S_n – Soil moisture sensor, H_n – Humidity sensor, R_n – Rain sensor, W_n – Wind Sensor, and the values of every sensor from 1… n.

If location number is one then the set of L_1 is, L_1= {t_1, s_1, h_1, r_1, w_1}. Likewise, L_2, L_3, L_4 and L_5 sets are defined. In the suggested method, 5 diverse locations are assumed, hence the overall count of sensors in every category shall be represented as,

T= {t_1, t_2, t_3, t_4, t_5},
S ={s_1, s_2, s_3, s_4, s_5},
H = {h_1, h_2, h_3, h_4, h_5},
R = {r_1, r_2, r_3, r_4 r_5}
W = {w_1, w_2, w_3, w_4, w_5}

Hence, L is represented as L = {T, S, H, R, W}

CID method for noise identification and elimination

for (int i = 0; i ≤ 25; i+=2) **// One observation every two hour**

collect r_1(Td[i])

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

if(r_1 (Td[i]) < r_1(Td[i+1])) **//Checking Redundant values based on TimeTd**

remove r_1(Td[i])

compute rest of R, and all elements in T,W, H, S

if(compare r_1 with R (μ), R(M), & R(Z) **// Checking point noise and error value**

*replace with R(μ), R(M), & R(Z)***// Common for rest of R and T, W, S, H**

if (r_1> 0) **// M,Z,μ are Calculated with respective to Td[i] value**

compute rest of R, and all elements in T,W, H, S

else

compute rest of R, and all elements in T,W, H, S

end if

end for

Section IV: Result and discussion

This section examines the suggested CID Method's effectiveness utilizing standard metrics like precision, F1 score, recall, and accuracy. Table 3 summarizes the information gathered. Lastly, the cleansed data is fed into a SVM classifier to evaluate the suggested CID method's efficiency.

On the acquired datasets, known pre-processing techniques like Iterative Class Noise Filter (INFFC), Meta Learner (MTL), and CEEDMAN are used and provided to the classifiers following cleansing. On the basis of performance measures, the suggested CID Method is then contrasted to current approaches. The suggested CID approach outperforms the competition in terms of accuracy, as demonstrated in Figure 1.

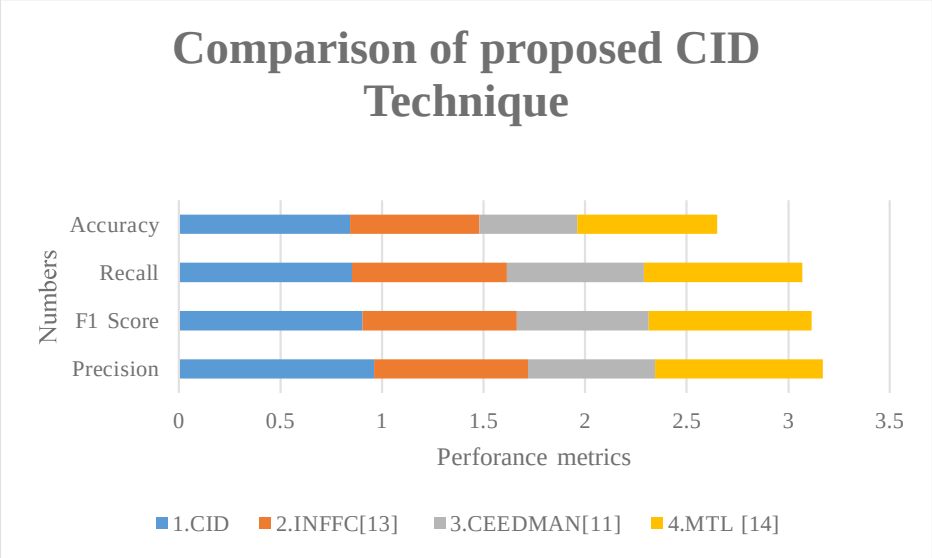


Figure 1: Comparison Result

Section V: Conclusion

For effective decision-making in the IoT, agricultural data must be pre-processed. The inconsistencies in the raw dataset obtained from the IoT ecosystem have an impact on decision-making efficiency and reliability. As a result, data refining is required. The suggested CID effectively manages noisy data. It is made up of three layers. The first layer gathers information from sensors located in the different areas, the second layer stores the gathered information, and the third layer cleans the information. The suggested method identifies noisy data and substitutes it with new data based on pre-defined parameters and central tendency metrics. Lastly, the findings were compared to current approaches, and the new strategy surpassed the competition by increasing accuracy of classification. Missing data and outliers might well be taken into account in the future to enhance accuracy.

References:

[1] Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W," A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", Sensors, 19(20), 4536, 2019, doi: 10.3390/s19204536.

[2] Assahli, S., Berrada, M., & Chenouni, D,"Data pre-processing from Internet of Things: Comparative study", Wireless Technologies, Embedded and Intelligent Systems (WITS), 2017.

[03] Morais, C. M. de, Sadok, D., &Kelner, J, "An IoT sensor and scenario survey for data researchers", Journal of the Brazilian Computer Society, doi:10.1186/s13173-019-0085-7,2019.

[04]Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W. "A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", Sensors, 19(20), 4536. doi:10.3390/s19204536, 2019.

[05] García-Gil, D., Luengo, J., García, S., & Herrera, F.,"Enabling Smart Data: Noise filtering in Big Data classification", Information Sciences. doi:10.1016/j.ins.2018.12.002, 2018.

[06] Peter Wlodarczak, Mustafa Ally, Jeffrey Soar, "Data Mining in IoT", In Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17) , ISBN 978-1-4503-4951, https://doi.org/10.1145/3106426.3115866, 2017.

[07] Lan, K., Fong, S., Song, W., Vasilakos, A., &Millham, R, " Self-Adaptive Pre-Processing Methodology for Big Data Stream Mining in Internet of Things Environmental Sensor Monitoring", Symmetry, 9(10), 244, doi:10.3390/sym9100244, 2017.

[08] Jamshed, Huma& Khan, M. &Khurram, Muhammad & Inayatullah, Syed &Athar, Sameen, "Data Preprocessing: A preliminary step for web data mining". 206-221, 2015, Doi: 10.17993/3ctecno.2019.specialissue2.206-221, 2019.

[09] Asiya Sulthana ,Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOT Enabled Telecardiology System", International Journal of Engineering & Technology, 7 (2.17) (2018) 74-78, 2018.

[10] Liu, Y., Dillon, T., Yu, W., Rahayu, W., &Mostafa, F, "Noise removal in the presence of significant anomalies for Industrial IoT sensor data in manufacturing", IEEE Internet of Things Journal, 1–1. doi:10.1109/jiot.2020.2981476, 2020.

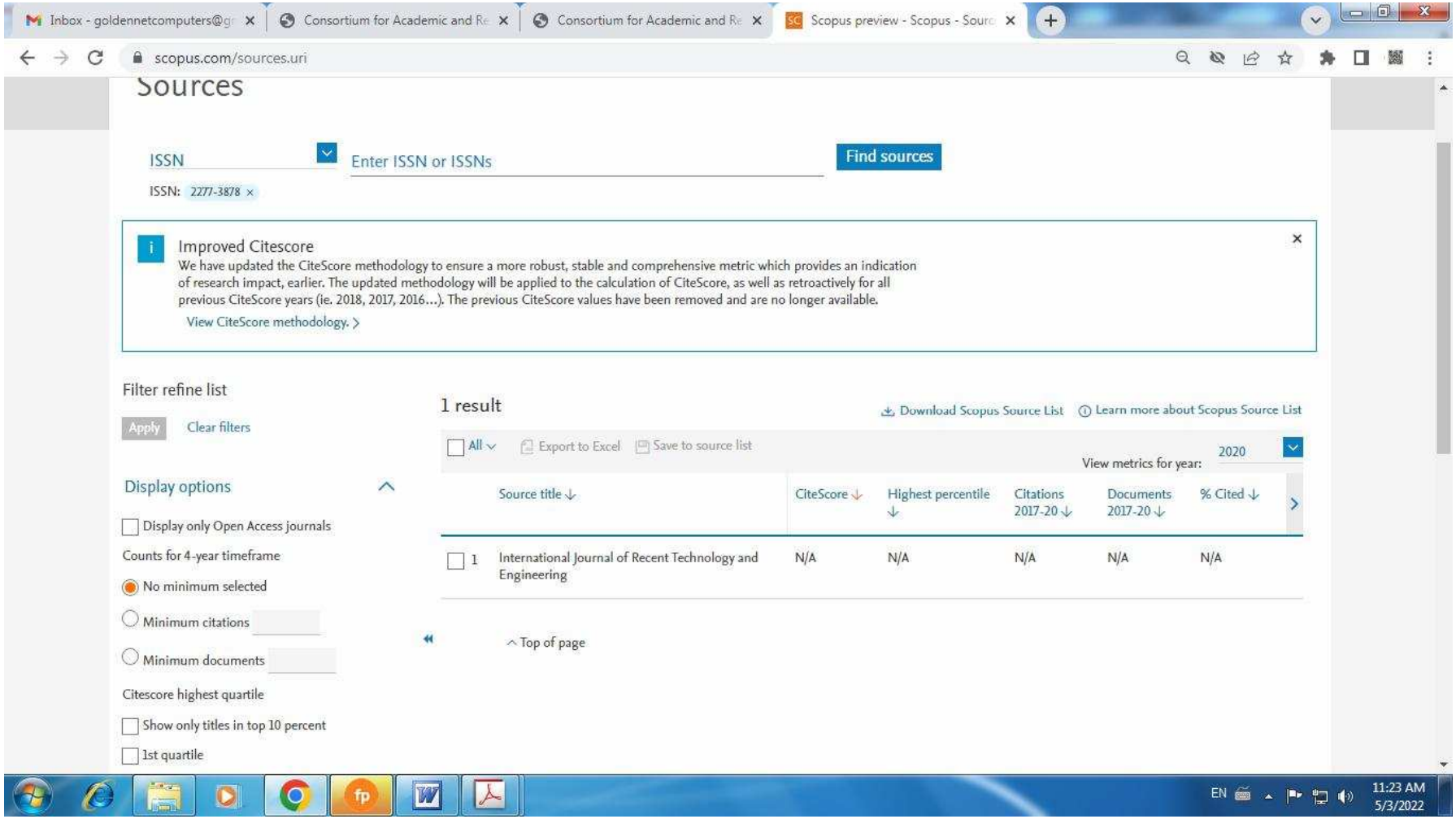
[11] Wang, Jianzhou; Wang, Ying; Li, Zhiwu; Li, Hongmin; Yang, Hufang, "A combined framework based on data preprocessing, neural networks and multi-tracker optimizer for wind speed prediction", Sustainable Energy Technologies and Assessments, 40, 100757–. doi:10.1016/j.seta.2020.100757, 2020.

[12] Sanyal, Sunny; Zhang, Puning," Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications", IEEE Access, Vol.6, 67830–67840, , doi:10.1109/ACCESS.2018.2878640, 2018.

[13] Sáez, J. A., Galar, M., Luengo, J. & Herrera, F., "INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control", Information Fusion, 27, 19–32, 2016.

[14] Garcia, L. P., de Carvalho, A. C. & Lorena, A. C. 2016a. "Noise detection in the meta-learning level. Neurocomputing, 176, 14–25, 2016.

[15] Choh Man Teng, **"**A Comparison of Noise Handling Techniques",FLAIRS-01 Proceedings, 2002.


Inbox - goldennetcomputers@g
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Sources
ISSN
Enter ISSN or ISSNs
Find sources
ISSN: 2277-3878
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
Citescore highest quartile
Show only titles in top 10 percent
1st quartile
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
International Journal of Recent Technology and Engineering
N/A
N/A
N/A
N/A
N/A
Top of page
EN
11:23 AM
5/3/2022

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.

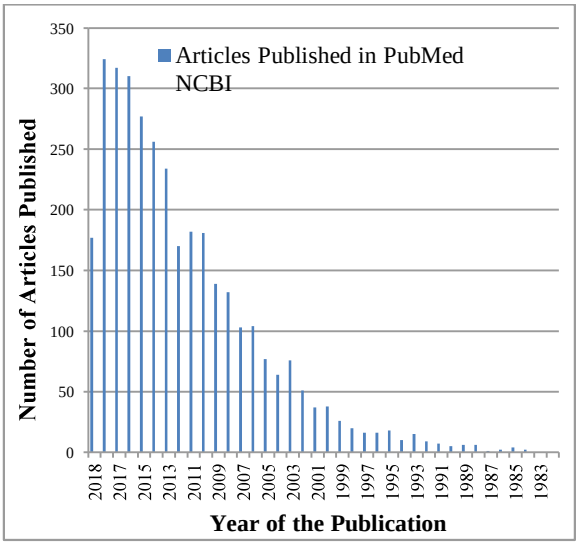


Fig 1: Number of publications index by Pub Med with keywords "Predictive and Diabetes." The 2018 count is extrapolated based on the number from June 27, 2018

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.

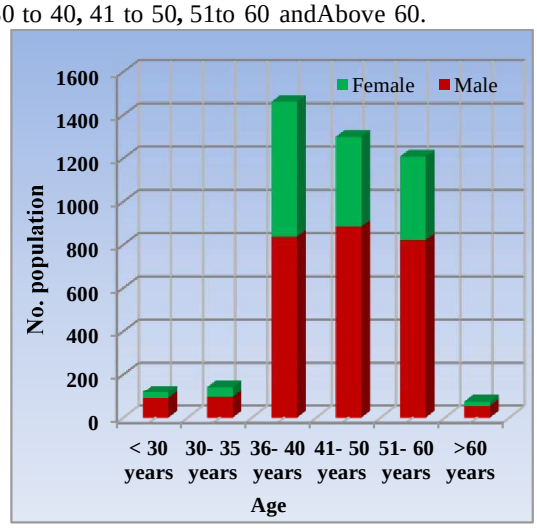


Fig 2: Distribution of Age and Sex

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.

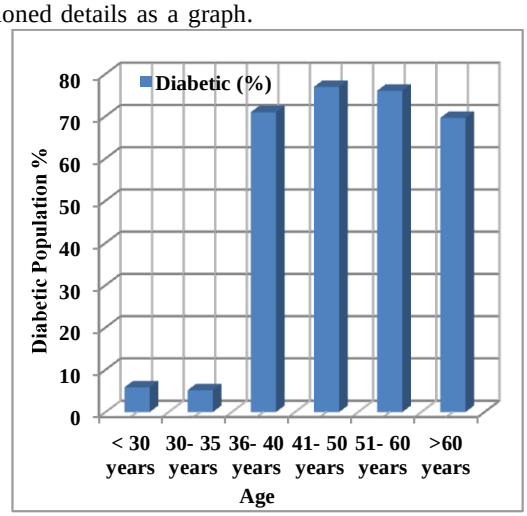


Fig 3: Diabetic and Age

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.

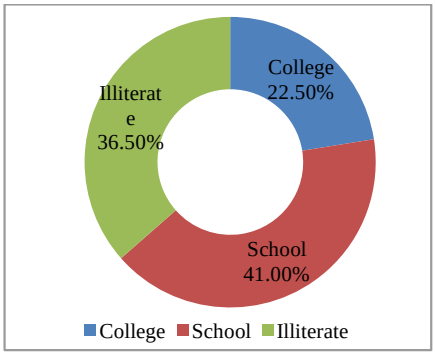


Fig 4: Education and Diabetic

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.

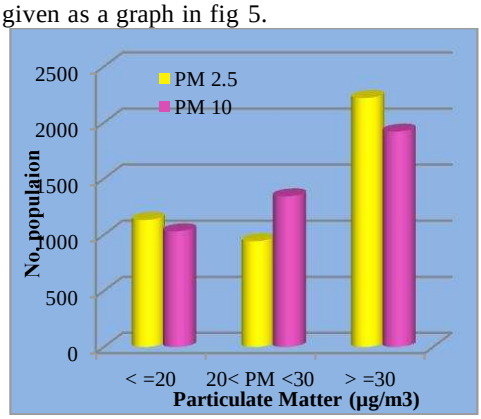


Fig 5: Air Pollution and Diabetic

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.

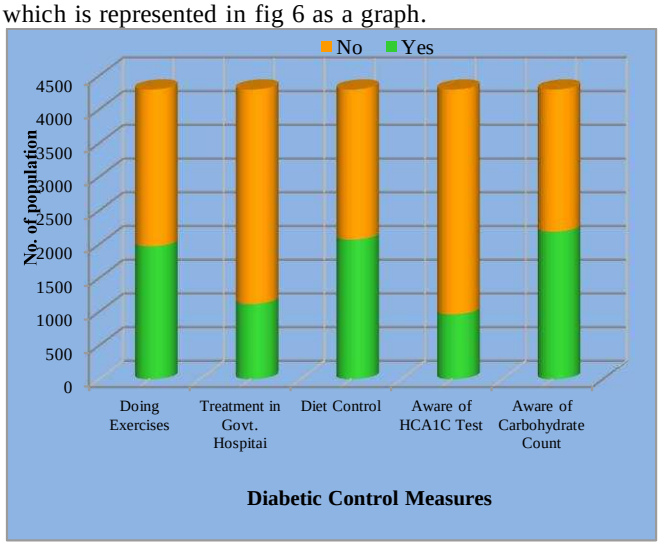


Fig 6: Diabetic Control measure

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.

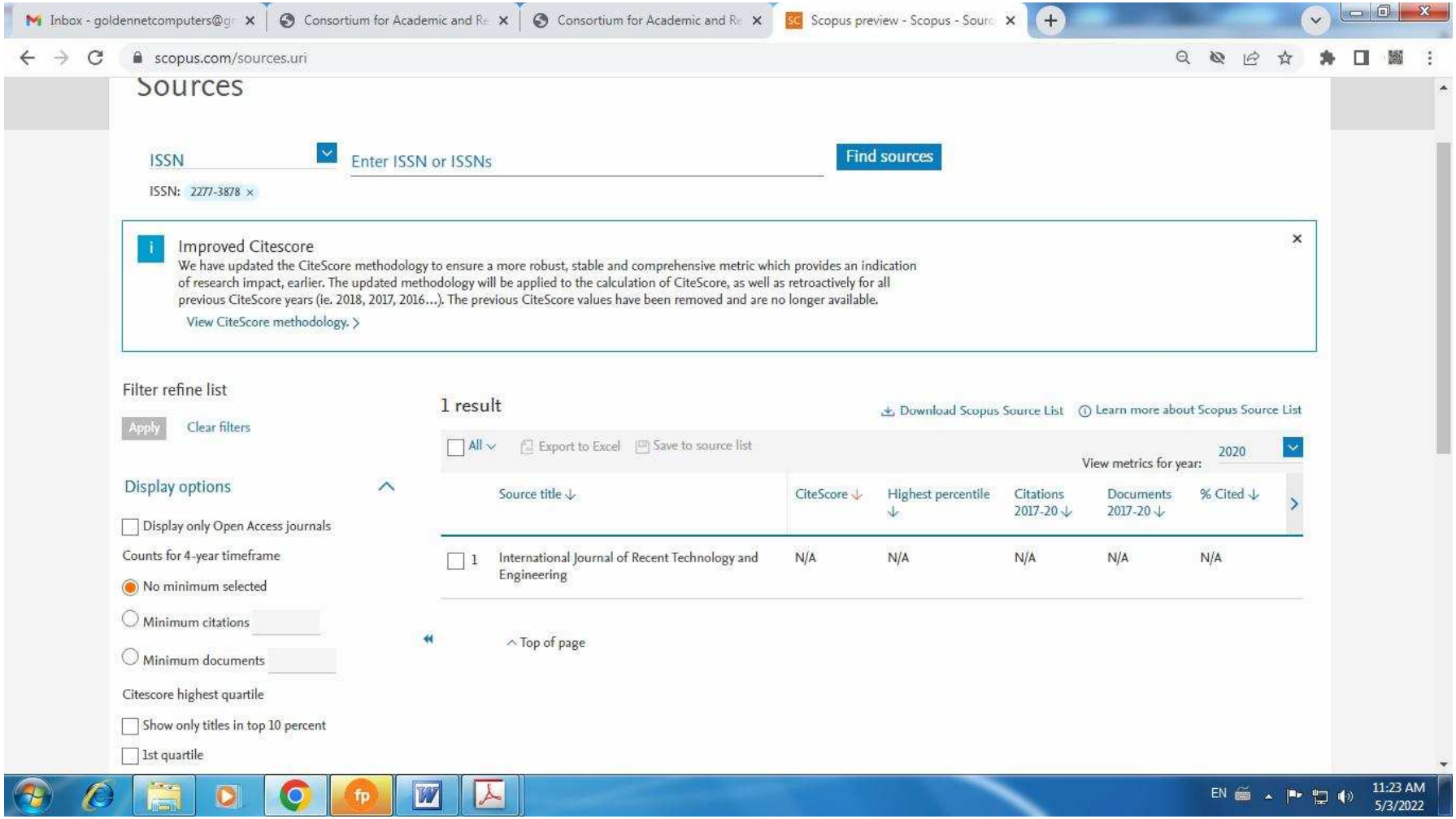
AUTHORS PROFILE

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


Inbox - goldennetcomputers@g
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Sources
ISSN
Enter ISSN or ISSNs
Find sources
ISSN: 2277-3878
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
Citescore highest quartile
Show only titles in top 10 percent
1st quartile
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
International Journal of Recent Technology and Engineering
N/A
N/A
N/A
N/A
N/A
Top of page
EN
11:23 AM
5/3/2022

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.

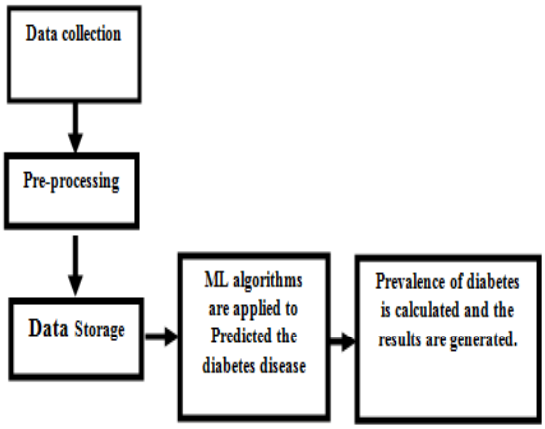


Fig 1: Work flow of proposed methodology

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.

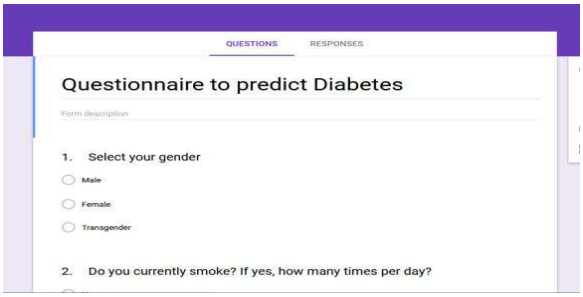


Fig 2: Questionnaire model based data collection

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.

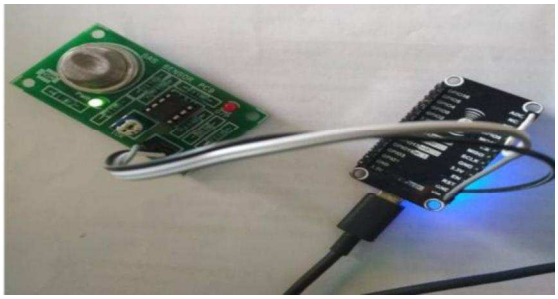


Fig 3: Data collection from sensor

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.

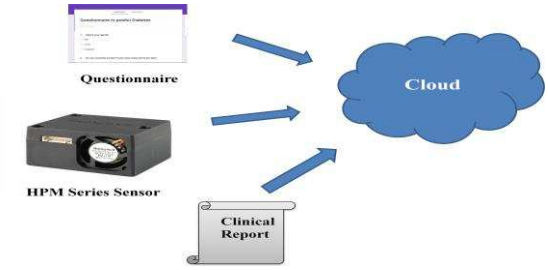


Fig 4: Collection of various data)

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.

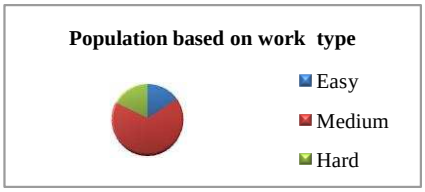


Fig 6: Population divided by work type.

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.

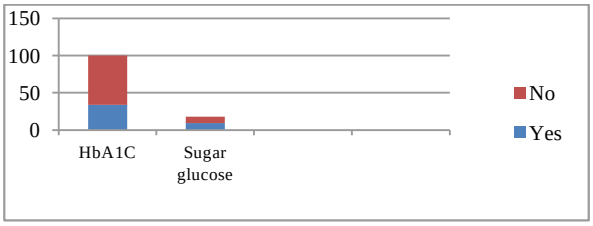


Fig 7: Awareness about Diabetes Mellitus

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.

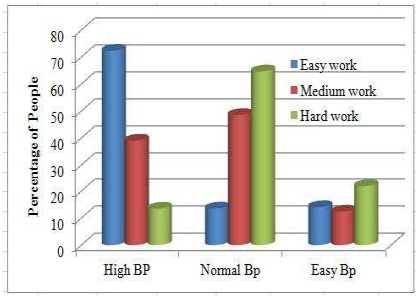


Fig 8: Work type vs Blood pressure level

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.

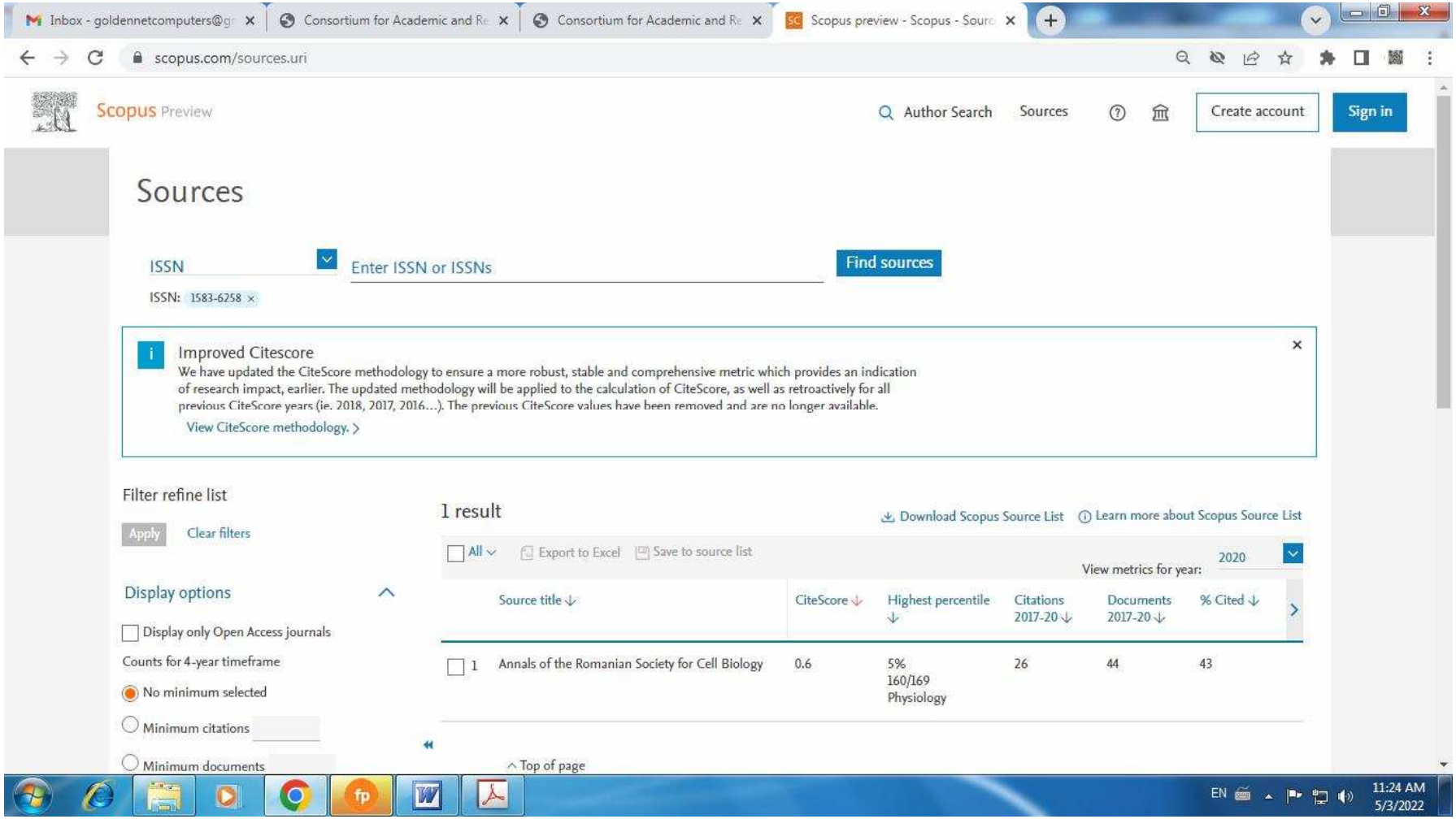
AUTHORS PROFILE

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


Inbox - goldennetcomputers@g
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Scopus Preview
Author Search
Sources
Create account
Sign in
Sources
ISSN
Enter ISSN or ISSNs
Find sources
ISSN: 1583-6258
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
Annals of the Romanian Society for Cell Biology
0.6
5%
160/169
Physiology
26
44
43
Top of page
EN
11:24 AM
5/3/2022

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

DaRoN: A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency

V. A. Jane[1], Dr. L. Arockiam[2],
[1,2] Department of Computer Science, St. Joseph's College Tiruchirappalli.

Abstract

The Internet of Things (IoT) is a significant technology that offers well-organized and trustworthy solutions for the innovation of many domains. Agriculture is one of the most concerned fields in IoT, where IoT based solutions are used to automate the maintenance and monitoring process with least human intervention. Large scale IoT based agricultural environment generates a large amount of data every moment. The agro-production environment is complex and there are numerous discrepancies in the collected raw data that cannot be directly traced by analysis and mining. To handle these inconsistencies in IoT agricultural data, this paper proposes a technique called **Detection and Removal of Noise(DaRoN)**. The proposed technique removes the null values, error values, repeated values, incomplete values, and irrelevant values using measures of central tendency. In addition, a comparative analysis was performed with existingnoise removal techniques and the performance is measured using the Support Vector Machine(SVM) classifier. In this proposed research work, noisy data is eliminated to enhance classification accuracy. The DaRoN technique will be useful for improving the quality of collected data in agricultural environment.

Key Words

Noise, Data cleaning, IoT, Preprocessing, Noise removal, Smart Agriculture.

Section I: Introduction

IoT is a predominant technology which makes many applications smarter using its features [1]. In the past, gathering data in agriculture environment was a difficult task especially in monitoring systems but IoT removes all those strenuous part with the help of sensors. Here, sensors play a vital role in data collection and generates enormous data every day. These data contain missing values, noise, outliers, and duplicate values [2]. If any one of the above is present in the collected data, then it will reduce the quality of outputs. Among which, Noise is one of the most considerable one and it is defined as meaningless information like, corrupted values, repeated values, error values, null values etc., These problems occur due to the reasons such as connection error, detection error, and collision problem in IoT [3]. If the dataset contains noisy values, then many problems will occur during the analytical process.

Noise is classified into two types such as point noise and continuous noise. The Point noise has sudden deviation from other data points.So this could be identified easily. The Continuous noise is difficult to identify because the deviation gets increased from point to points. For removing these types of noise, mean, median and mode methods are used. Noise can also be categorized based on the occurrences in the dataset. If the noise occurs in the class column then it is called *class noise*. If the noise occurs in the attribute column then it is called *attribute noise*. In contrast to *class noise*, *attribute noise* is more harmful because it directly affects the data. Thus, noise in the dataset will affect the accuracy of the analytics model [4]. So, there is a need for data pre-processing.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

Pre-processing techniques [5] are categorized into various types such as data cleaning, data integration, data transformation and data reduction. This paper focuses on noise removal and it comes under the process of data cleaning.The rest of the paper explains more details about the proposed technique and the paper is organized as follows, section II explores the related works on the relevant area, Section III describes the methodology of the proposed work, section IV summarizes the results & discussion and section V concludes the work.

Section II: Related Work

Peter et al.,[6] overviewed the role of Data mining in IoT. This work discussed about all the technologies, methods and algorithms related to the Data mining process with respect to various IoT applications. Also, it described the role of data management in smart environments.

Kun et al., [7] proposed clustering-based particle swarm optimization (CPSO) approach to handle data in the DSM (Data Stream Mining). In which, sliding window technique was used for data segmentation and SFX (Statistical Feature Extraction) was used for variable partitioning. The proposed approach was implemented using five types of IoT data set (Home, Gas, Ocean, and Electricity). The results were evaluated, and the proposed approach improved the accuracy but increased the complexity of algorithms and the over fitting problem.

HumaJamshed et al.,[8] discussed about various pre-processing techniques for mining and analysis tasks. In this work, the important methods of data pre-processing were described which includes data cleaning, data transformation, data reduction and data integration. The author proposed a technique for the same and explained with simple text data case study. The proposed technique dealt with noise removal, tokenization, and normalization. The paper concluded that the advanced techniques like machine learning improved the effectiveness of pre-processing.

Asiya et al., [9] compared the performance of noise cancellation techniques in IoT enabled Telecardiology System. The techniques which were taken for comparison were LMS (Least Mean Square), NLMS (Normalized Least Mean Square), CLLMS (Circular Leaky Least Mean Square) and VSS-CLLMS (Variable Step Size CLLMS). Baseline Wander (BW) elimination (lowest frequency in ECG (ElectroCardioGram)). VSS-CLLMS method achieved high SNRI (Signal to Noise Ratio Improvement). Theauthors focused on ECG data pre-processing with filtering mechanisms.

Liu et al., [10] proposed a technique to handle noise in IoT data by using anomaly detection technique. The proposed technique measured the rate of change and deviation by using a sliding window and statistical techniques. Also, it identified the noise in the dataset based on neighbour behaviour and erroneous data removal process was difficult if error was identified in the continuous neighbour. Here, the identification process consumed more time.

Wang et al.,[11] proposed a framework for wind data pre-processing and prediction. In this proposed work,Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEDMAN)technique used to remove noise in the wind data and MTO (multi-tracker optimizer) was used for error detection. Finally, neural network layers were utilized for model building. The proposed CEEDMDAN technique was suitable only for limited sized datasets and while large datasets were considered it increased the mean error.

Sanyall et al.,[12] proposed a scheme to handle the veracity problems (Noise, Missing values, Outliers and redundancy) in IoT sensor data. The proposed scheme consisted of two parts, first part dealt with data aggregation using cluster method and the second part dealt with data pre-processing using robust dominant subspace estimation and tracking methods. Random

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

outputs generated by dominant subspace selection increased outliers so the overall performance was decreased.

Sáez et al.,[13] presented a method Iterative Class Noise Filter (INFFC) that combined many classifiers for detecting noise in an iterative manner.The filtration method was introduced to identify the noise by eliminating the process of noise detection at each new iteration.

Garcia et al.,[14] improved noise detection using an ensemble of noise filtering methods. The proposed approach Meta Learner (MTL) reduced the redundant data in the dataset, as well as eliminated the irrelevant data. For that, Meta features were created from corrupted datasets and provided a meta-learning model that predictednoisy data.

Section III: DaRoN Proposed Technique

In the agricultural scenario, irrigation system requires constant monitoring without human intervention. To automate this process, the proposed DaRoN technique uses IoT sensors to collect data and stores the collected data in cloud. Later the collected data are pre-processed using the measures of central tendency and the performance of the pre-processed dataset is analysed using Support Vector Machine (SVM) classifier. In traditional noise handling techniques, there are 3 phases namely,Robust (detection of any analysis errors to make the data standardized), Filtering (using various measures to remove noise), and Polishing (Replacing error values)[15]. The novelty of the proposed DaRoN is that it combines the 3 phases into one to produce a Noiseless dataset.

Robust and filtering are done using pre-defined conditions and polishing is done by using measures of central tendency. The work flow of the DaRoN technique is given below in figure 1.

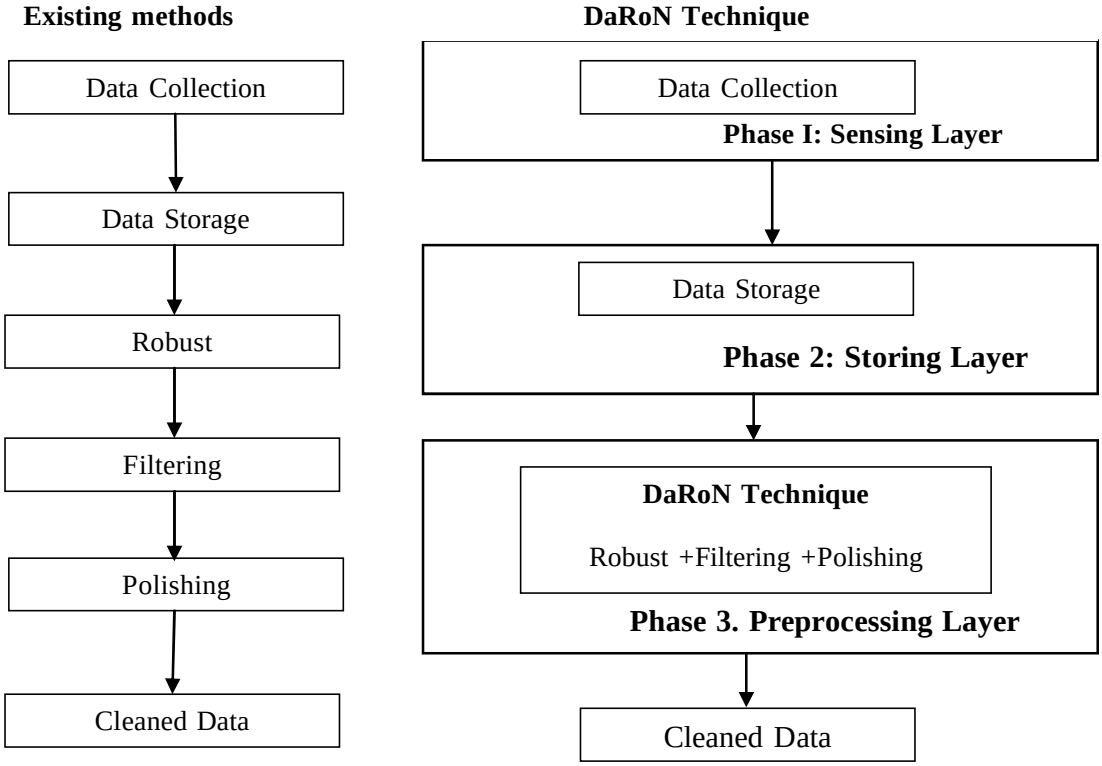


Figure 1: Work flow of the Existing vs. DaRoN technique

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

From the workflow of the proposed work, the pre-processed agricultural dataset yields noiseless cleaned dataset which increases the performance of the classifier.

Phase I: Sensing Layer

Phase one deals with the data collection that is done by using various IoT sensors in agricultural environment. There are five sensors used namely, humidity sensor, temperature sensor, soil moisture sensor, wind speed sensor and rain sensor. Sensors are placed in different places and connected to the cloud. Each sensor plays a unique role in monitoring the environment and continuously collects data. Humidity sensor collects the data about moisture level in the air. This data will be useful in determining whether irrigation is needed or not. Soil moisture sensor measures the percentage of water present in the soil. In this work, both the humidity and soil moisture sensor data are taken together to make irrigation decisions. The temperature sensor is generally used to measure temperature level from time to time. The rain sensor is used to collect rain level. The primary work of this sensor is to shut down the entire irrigation system during heavy rain fall. Data from these sensors are collected automatically and sent to the server directly for further processing.

Phase 2: Storing Layer

Second layer is storing layer, which is used for data storage purpose. Basically, data can be stored on local devices, but to handle large data, cloud storage is the best. So, the proposed technique uses cloud to store the data. Many open-source clouds are available, one of them is ThinkSpeak cloud server which provides open-source computing model, where data can be stored and retrieved remotely with the help of internet. The stored data is maintained, operated, and managed by a service provider. In ThinkSpeak, an account is created and built with various fields such as soil field, humidity field, temperature field, and rain field to store their respective information. After that, the stored data is forwarded to the preprocessing layer.

Phase 3. Preprocessing Layer

In this layer, the proposednoise removal technique is used.This novel technique uses the measures of central tendency. Traditional Noise removal techniques use three phasessuch as robust, filtering and polishing.But, the proposed DaRoN technique combines these three phases in a single phase by using the measures of Central tendency which gives better performance. The proposed technique selects thenearestmean value to replace repetitive and null values. Nearest Mode value is selected to remove Point Noise.All replacementsare done with respect to Time Details (Td).The Central tendency measures are listed below.

$$\text{Mean } (\mu) = \frac{\text{sum of all elements}}{\text{Total number of Elements}}$$

$$\text{Median (M)} = L + h\frac{((fm - f1))}{((fm - f1) - (fm - f2))}$$

$$\text{Mode (Z)} = \frac{(n+1)}{2}$$

Let $L = \{L_1, L_2 \ldots L_n\}$, where, L_1, L_2 …$L_n$are different locations.

Each location has various sensors that are T_n, S_n, H_n, R_n and W_n where n denotes number oflocations, T_n – Temperature sensor, S_n – Soil moisture sensor, H_n – Humidity sensor, R_n – Rain sensor, W_n – Wind Sensor, and the values of each sensor from 1… n.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

If location number is one then the set of L_1 is, L_1= {t_1, s_1, h_1, r_1, w_1}. Similarly, L_2, L_3, L_4 and L_5 sets are defined. In the proposed work, 5 different locations are considered, so the total number of sensors in each category can be written as,

T= {t_1, t_2, t_3, t_4, t_5},
S ={s_1, s_2, s_3, s_4, s_5},
H = {h_1, h_2, h_3, h_4, h_5},
R = {r_1, r_2, r_3, r_4 r_5}
W = {w_1, w_2, w_3, w_4, w_5}

Therefore, L can be written as L = {T, S, H, R, W}

DaRoN Technique for noise detection and removal

for (int i = 0; i ≤ 25; i+=2) **// One observation per two hour**

collect r_1(Td[i])

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

if(r_1 (Td[i]) < r_1(Td[i+1])) **//Checking Redundant values based on TimeTd**

remove r_1(Td[i])

compute rest of R, and all elements in T,W, H, S

if(compare r_1 with R (μ), R(M), & R(Z) **// Checking point noise and error value**

replace with R(μ), R(M), & R(Z) **// Common for rest of R and T, W, S, H**

if (r_1> 0) **// M,Z,μ are Calculated with respective to Td[i] value**

compute rest of R, and all elements in T,W, H, S

else

compute rest of R, and all elements in T,W, H, S

end if

end for

Section IV: Result and discussion

This section describes the performance of the proposed DaRoN Technique using the conventional measures such as precision, F1 score, recall and accuracy.Table 3 shows the details of the collected data. Finally, the cleaned dataset is applied to the SVM classifier for analysing the performance of the proposed DaRoN technique.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

Table 3 Collected Data Details

Type				**Amount**			
Total Data (Rows)				19,520(6 Months Data)			
Noise				9,760			
Point Noise	Repetitive	Collision	Null	4782	4326	118	652
Features(Columns)				32			

The existing pre-processing methods such as Iterative Class Noise Filter (INFFC), Meta Learner (MTL) and (CEEDMAN) are applied on the collected dataset and fed to the classifier after cleaning. Then the proposed DaRoN Technique is compared with existing techniques based on the performance metrics. The proposed DaRoN technique enhances the accuracy than others and the comparison results are shown in Figure 2.

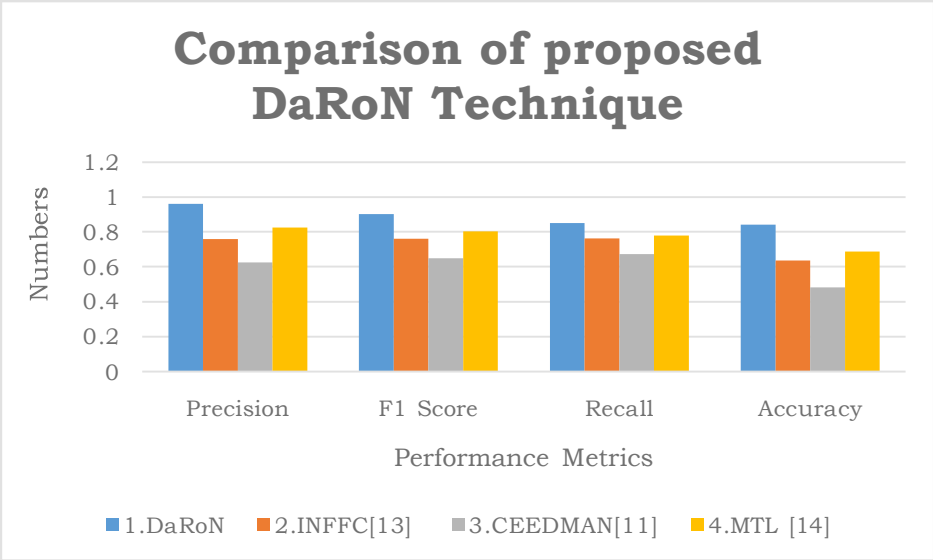


Figure 2:Comparison Result

Section V: Conclusion

In IoT agricultural data needs pre-processing for efficient decision making. The raw data collected from IoT environment has inconsistency issues which affect the efficiency and accuracy of decision making. So, refinement of data is needed. The proposed DaRoN handles the noisy data efficiently. It consists of three layers. First layer collects data from sensors placed in various locations, the collected data is stored in the second layer, and the third layer performs data cleaning process. The proposed technique detects noisy data and replaces it based on pre-defined conditions and measures of central tendency. Finally, the results were compared with existing methods and the proposed technique outperformed others by improvingthe classification accuracy. In future, issues like missing values and outliers may also be considered to further improve accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

References:

[1] Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W," A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", *Sensors*, 19(20), 4536, 2019, doi: 10.3390/s19204536.

[2] Assahli, S., Berrada, M., &Chenouni, D,"Data pre-processing from Internet of Things: Comparative study", *Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2017.

[03] Morais, C. M. de, Sadok, D., &Kelner, J, "An IoT sensor and scenario survey for data researchers*", Journal of the Brazilian Computer Society*, doi:10.1186/s13173-019-0085-7,2019.

[04]Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W. "A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", *Sensors*, 19(20), 4536. doi:10.3390/s19204536, 2019.

[05] García-Gil, D., Luengo, J., García, S., & Herrera, F.,"Enabling Smart Data: Noise filtering in Big Data classification", *Information Sciences*. doi:10.1016/j.ins.2018.12.002, 2018.

[06] Peter Wlodarczak, Mustafa Ally, Jeffrey Soar, "Data Mining in IoT", *In Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17)* , ISBN 978-1-4503-4951, https://doi.org/10.1145/3106426.3115866, 2017.

[07] Lan, K., Fong, S., Song, W., Vasilakos, A., &Millham, R, " Self-Adaptive Pre-Processing Methodology for Big Data Stream Mining in Internet of Things Environmental Sensor Monitoring", *Symmetry*, 9(10), 244, doi:10.3390/sym9100244, 2017.

[08] Jamshed, Huma& Khan, M. &Khurram, Muhammad &Inayatullah, Syed &Athar, Sameen, "Data Preprocessing: A preliminary step for web data mining". 206-221, 2015, Doi: 10.17993/3ctecno.2019.specialissue2.206-221, 2019.

[09] AsiyaSulthana ,Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOTEnabled Telecardiology System", *International Journal of Engineering & Technology*, 7 (2.17) (2018) 74-78, 2018.

[10] Liu, Y., Dillon, T., Yu, W., Rahayu, W., &Mostafa, F, "Noise removal in the presence of significant anomalies for Industrial IoT sensor data in manufacturing", *IEEE Internet of Things Journal*, 1–1. doi:10.1109/jiot.2020.2981476, 2020.

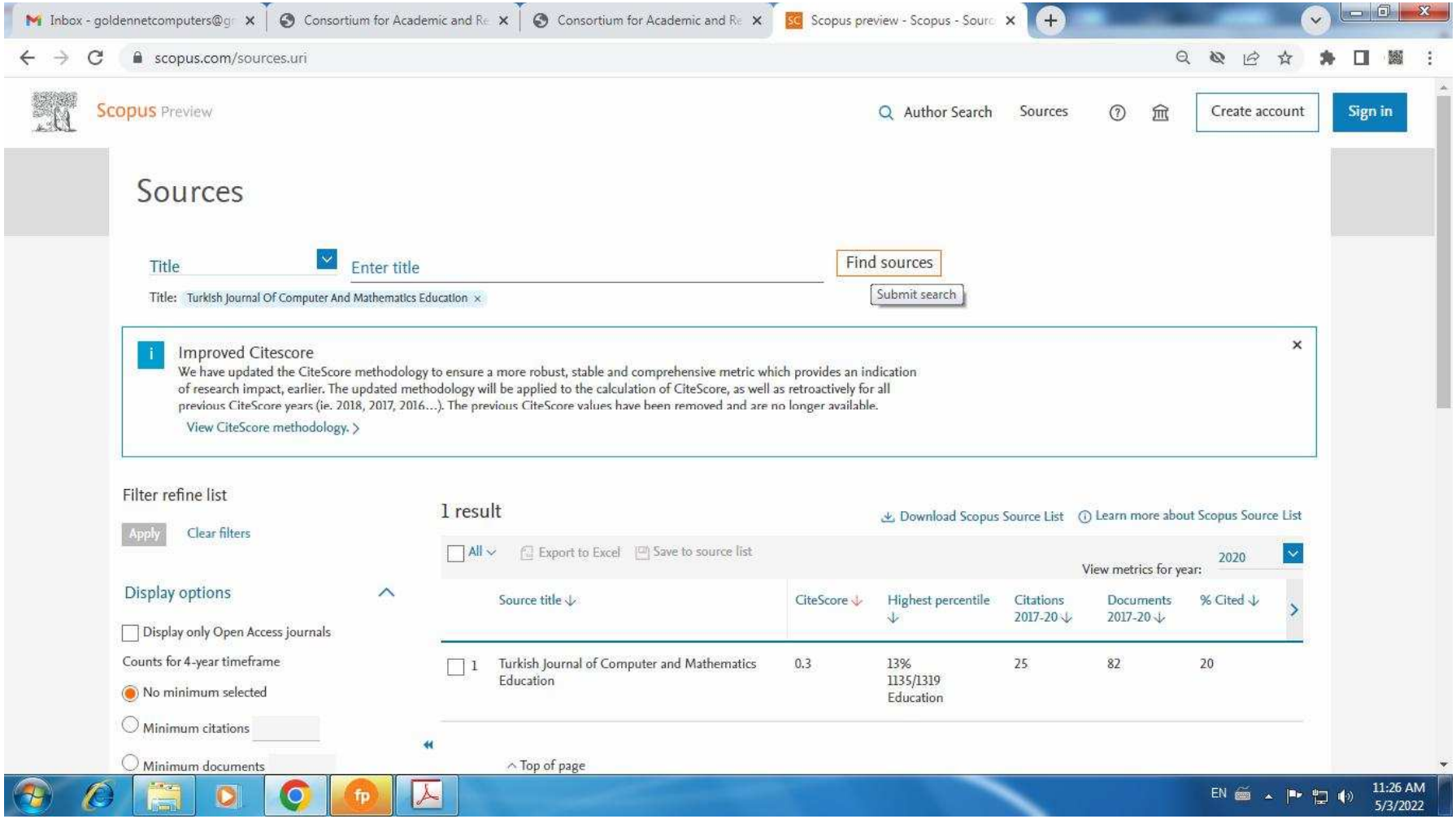
[11] Wang, Jianzhou; Wang, Ying; Li, Zhiwu; Li, Hongmin; Yang, Hufang, "A combined framework based on data preprocessing, neural networks and multi-tracker optimizer for wind speed prediction", *Sustainable Energy Technologies and Assessments*, 40, 100757–. doi:10.1016/j.seta.2020.100757, 2020.

[12] Sanyal, Sunny; Zhang, Puning," Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications", *IEEE Access*, Vol.6, 67830–67840, , doi:10.1109/ACCESS.2018.2878640, 2018.

[13] Sáez, J. A., Galar, M., Luengo, J. & Herrera, F., "INFFC: an iterative class noise filter based on the fusion ofclassifiers with noise sensitivity control", *Information Fusion*, 27, 19–32, 2016.

[14] Garcia, L. P., de Carvalho, A. C. & Lorena, A. C. 2016a. "Noise detection in the meta-learning level. *Neurocomputing*, 176, 14–25, 2016.

[15] Choh Man Teng, **"**A Comparison of Noise Handling Techniques",*FLAIRS-01 Proceedings*, 2002.


Inbox - goldennetcomputers@g
Consortium for Academic and Re
Consortium for Academic and Re
Scopus preview - Scopus - Sourc
scopus.com/sources.uri
Scopus Preview
Author Search
Sources
Create account
Sign in
Sources
Title
Enter title
Find sources
Submit search
Title: Turkish Journal Of Computer And Mathematics Education
Improved Citescore
We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.
View CiteScore methodology. >
Filter refine list
Apply
Clear filters
Display options
Display only Open Access journals
Counts for 4-year timeframe
No minimum selected
Minimum citations
Minimum documents
1 result
Download Scopus Source List
Learn more about Scopus Source List
All
Export to Excel
Save to source list
View metrics for year:
2020
Source title
CiteScore
Highest percentile
Citations 2017-20
Documents 2017-20
% Cited
1
Turkish Journal of Computer and Mathematics Education
0.3
13% 1135/1319 Education
25
82
20
Top of page
EN
11:26 AM
5/3/2022

Survey on IoT Data Preprocessing

V.A. Jane[a] and Dr. L. Arockiam[b]

[a] Department of computer Science, St. Joseph's College, Trichy, Tamilnadu, India 62001.
[b] Department of computer Science, St. Joseph's College, Trichy, Tamilnadu, India 62001.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Internet of Things (IoT) is a growing technology in all fields of science and engineering. The amount of data emitted by the sensors used in the various fields is high. Therefore, efficient knowledge from such large datasets is a clear requirement of many users. This large data is far from perfect; it has many defects (such as noise, missing values, outliers etc.) and is not suitable for analysis because it can lead to incorrect conclusions. So, data preprocessing is a required technique for such data. Data preprocessing is an important and essential step, the main goal of which is to dedicate techniques to clean, refine, repair and improve that raw data. This paper proposes a survey on IoT data preprocessing and its techniques. This paper discusses exiting research on data preprocessing in the IoT context and, introduces the background of IoT data preprocessing and present literature reviews of the advanced research on data preprocessing techniques. The classification of various preprocessing approaches with techniques is clearly depicted in the figure. Various approaches of preprocessing cleaning, transformation, reduction and integration are described. In addition, methods for such approaches in IoT data preprocessing are also discussed. IoT Data preprocessing techniques on various applications are tabulated. Finally, issues and challenges, most useful in future work, are discussed.

Keywords: IoT, Preprocessing, Data Cleaning, Noise handling

1. Introduction

Internet of Things basically refers to a network of objects that are connected to the Internet. It is an excellent computerization and analysis system across various industries such as agriculture, medical, transport, city, etc., [1]. Being connected to the Internet, one can collect data and send it over the internet, receive information from the internet, or do both. In the Internet of Things (IoT), the connected devices / sensors generate data enormously. These data are transferred to the cloud database for analyzing and to create smart applications. Data analytics is a very important technique to find insights from these data [2]. Before analyzing the data, data preprocessing plays a vital task owing to such kind of data with many defects like missing, noise, and inconsistent data. It is a kind of key stages in knowledge discovery process [3]. Low-quality data can undermine the effectiveness of successive learning algorithms. Therefore, avoiding the impact in quality, improves reliability of successive automated innovations and enhances decisions by taking appropriate preprocessing methods. There are various techniques involved in it, namely, data transformation, data reduction, data normalization, data cleaning, and data integration [5]. These techniques simplify the data by selecting or eradicating unnecessary features and dividing difficult constant feature spaces. During this process, the original input construction needs to be maintained and processing time need to be considered. Some benefits of data preprocessing are rapid training of learning methods, advanced generalization skills, as well as better understanding and easy interpretation of results [6]. This paper aims to survey on data preprocessing, its techniques and existing contributions of data preprocessing. This survey constructed as follows: In part II, The related works on data preprocessing in IoT environments and its techniques are discussed. Part III summarizes the techniques in various IoT based application, and part V concludes this work.

2. Related work

Hui et al., [7] reviewed the physical sensor errors that occur during the data-collection process. This paper described types of physical sensor errors, various error-detection mechanisms, error-correction techniques and also explained the differences between the techniques. Among error-detection and correction mechanisms, Principal Component Analysis (PCA) and Artificial Neural Network (ANN) provided better results.

Mathew et al., [8] compared various preprocessing techniques, namely, Kalman filter, z-scoring and moving Average filter. Firstly, preprocessing techniques were applied to the chemical sensor data to clean it. After that, the dataset is cleaned and evaluated using different classifiers such as Linear Discriminant Analysis (LDA), K Nearest Neighbor (KNN), and Support Vector Classifier (SVC). Finally, the performances of the various preprocessing techniques were calculated. Among these, it was observed that the Kalman filter technique provided better result than others.

Zena et al.,[9] reviewed methods of selecting and extracting features for high-dimensional Microarray cancer dataset. The writer discussed about the problems of irrelevant and redundant features in the micro array dataset. Also, the importance of dimensionality reduction, its advantages and drawbacks were discussed.

Chao et al.,[10] explained the process of data transmission in IoT environment. A preprocessing technique was adopted for reducing transmission time and increasing processing speedbut this work, only focused on reducing data transmission time.

Evgeniy [11] proposed architecture for preprocessing sensor data. Various preprocessing techniques suitable for proposed architecture were found. Streaming sensor data, the Univariate time series dataset, was utilized in this architecture.

Natarajasivan et al.,[12] proposed filter-based monitoring system for IoT Context. Sensors utilized in this work for sensing acceleration, position, vision, audio, temperature and direction. Kalman filter was utilized to process the collected data from those sensors and to evaluate the results using SVM. The proposed system consumed more time.

Cleber et al.,[13] surveyed all IoT application papers published since 2015. The author numbered the IoT application based on the usage. Smart home applications are used widely used by the researchers when compared with others. The sensor used in the smart environments is also discussed.

Rajalakshmi et al.,[14] discussed the function of IoT in smart appliances and summarized the problems such as data aggregation, scalability, data fusion, de-noising, heterogeneity, data outlier detection, real-time processing and missing data imputation. The author explained the usage of cloud, fog and edge computing in IoT to improve the analytics process and described the IoT data analytics process using a drone for traffic-monitoring system.

David et al.,[15] reviewed the data management problems in IoT environment, namely data collection, cleaning, integration, migration and processing. The author discussed the advanced data-processing technologies such as AI, machine learning, deep learning, and data mining.

Karinaer al.,[16] presented a survey on preprocessing techniques with relevant issues related to data mining. The fundamental concepts of data mining, preprocessing techniques and its issues were explained in detail. Moreover, it offered various solutions and discussed future directions.

García et al.,[17] proposed data preprocessing methods for big data era. The key areas of data preprocessing and current open challenges were explained. Moreover, the different data preprocessing techniques, namely, normalization, discretization, subset selection and extraction, feature indexers and encoders. In addition, other techniques for text mining were reviewed. Also, major issues of big data preprocessing were highlighted.

Jayaram et al.,[18] presented a study on data preprocessing methods. The main aim was to provide solutions for various problems of data preprocessing. The author focused data cleaning methods that includes filter, imputation, hybrid, wrapper and ensemble methods. The process and uses of each methods were described with examples. In particular, noise, data handling were considered and explanations regarding how to detect and treat it were given. Finally, the challenges while dealing with data cleaning at different fields were illustrated.

Huma Jamshed et al.,[19] discussed various big data Preprocessing techniques to clean data for further mining and analysis tasks. Initially, the important steps involved during the data preprocessing were explained. Then, a framework for web data preprocessing was proposed and each step was explained one by one. Finally, the simple text data was applied on the framework and preprocessing steps, like noisy removal, tokenization, normalization, were done.

Categories of preprocessing techniques

Data preprocessing is the process to make real world data more suitable for data mining process [20]. Real-world data is more noisy, contains missing values and a lot of ambiguous information, and these data are large in size. These factors cause the deterioration of the quality of the data during the result that obtaining after the mining or modeling. Therefore, before mining or modeling the data, it must be passed through improvement techniques known as data preprocessing. There are different techniques to perform such kind of process to make the data suitable for analyzing purposes. The categories of data pre-
processing techniques are shown in fig 1.

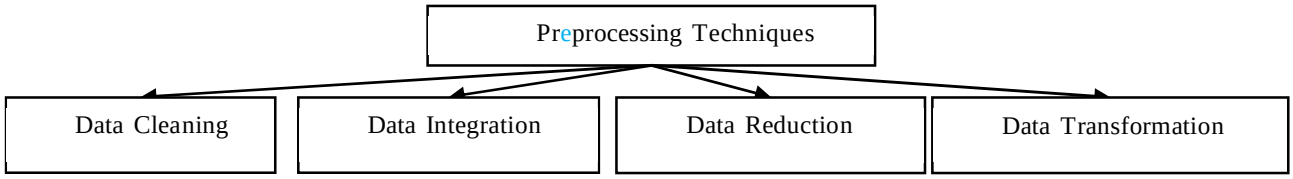


Fig: 1 Categories of data preprocessing Techniques

Data cleaning can be defined as the process of eliminating the erroneous and missing part in the data. The process of handling these noisy and missing values can be achieved by various ways, that shown in fig 2.

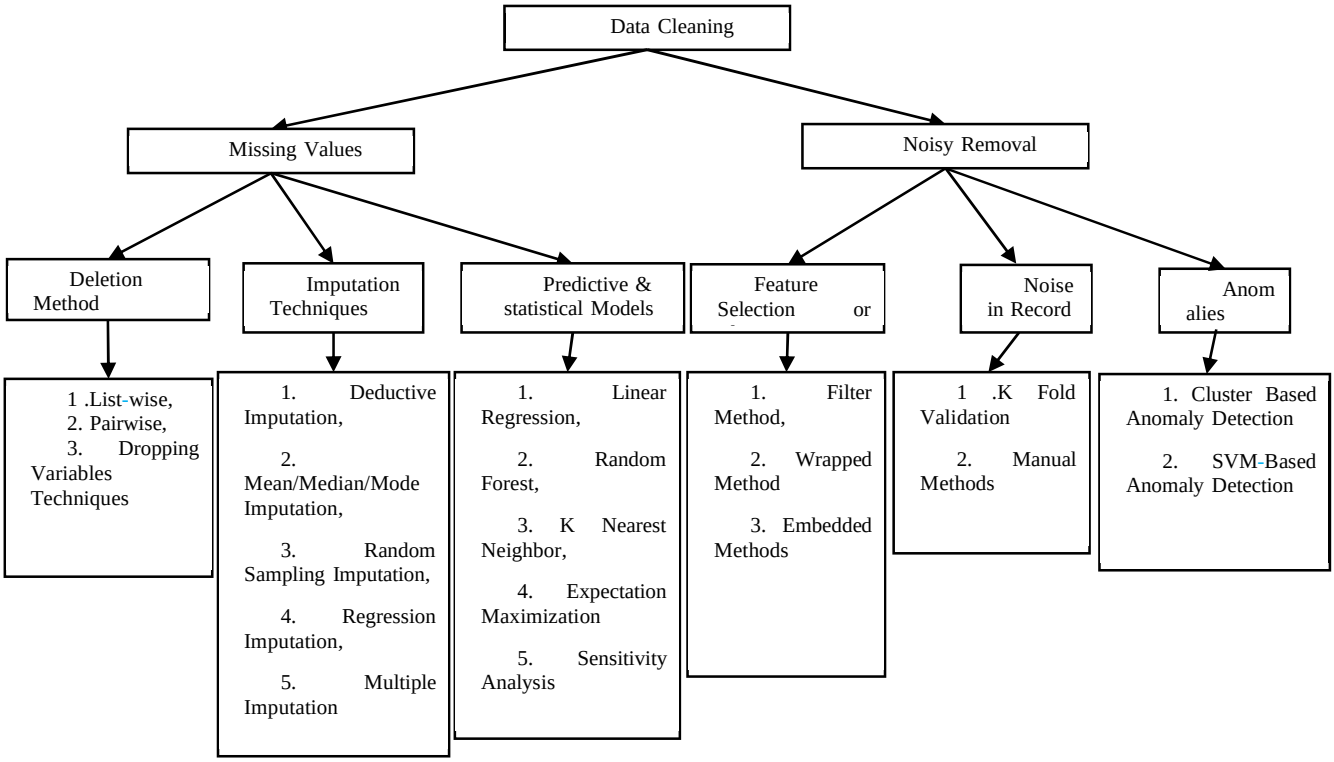


Fig: 2 Data Cleaning Techniques

Data integration is one important technique in preprocessing which combines data from different source and giving users an integrated view of this data. Mainly, Data integration is done through two main approaches, that are explains in the following fig 3.

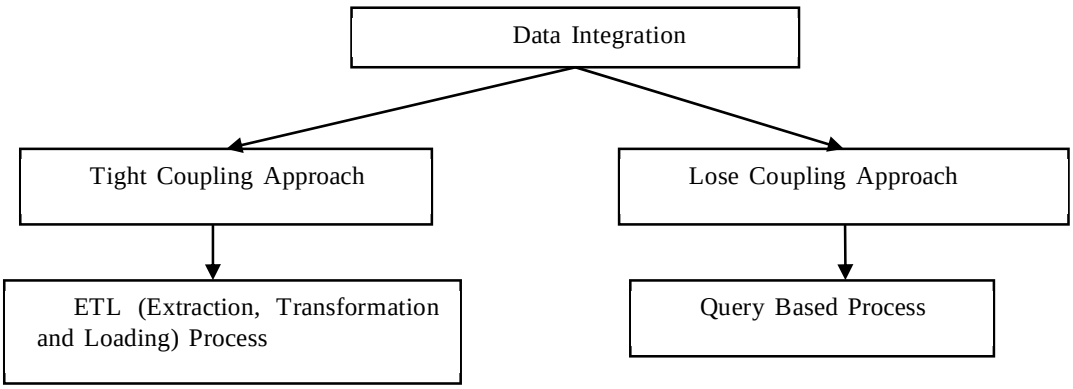


Fig: 3 Data Integration Techniques

Data reduction techniques can be used to obtain a data set, which are very small in size but yield, the same analytical results. Data reduction approaches utilized to diminish the unnecessary data as well as improve analytical process. Traditional, data reduction approaches are depicted in fig 4.

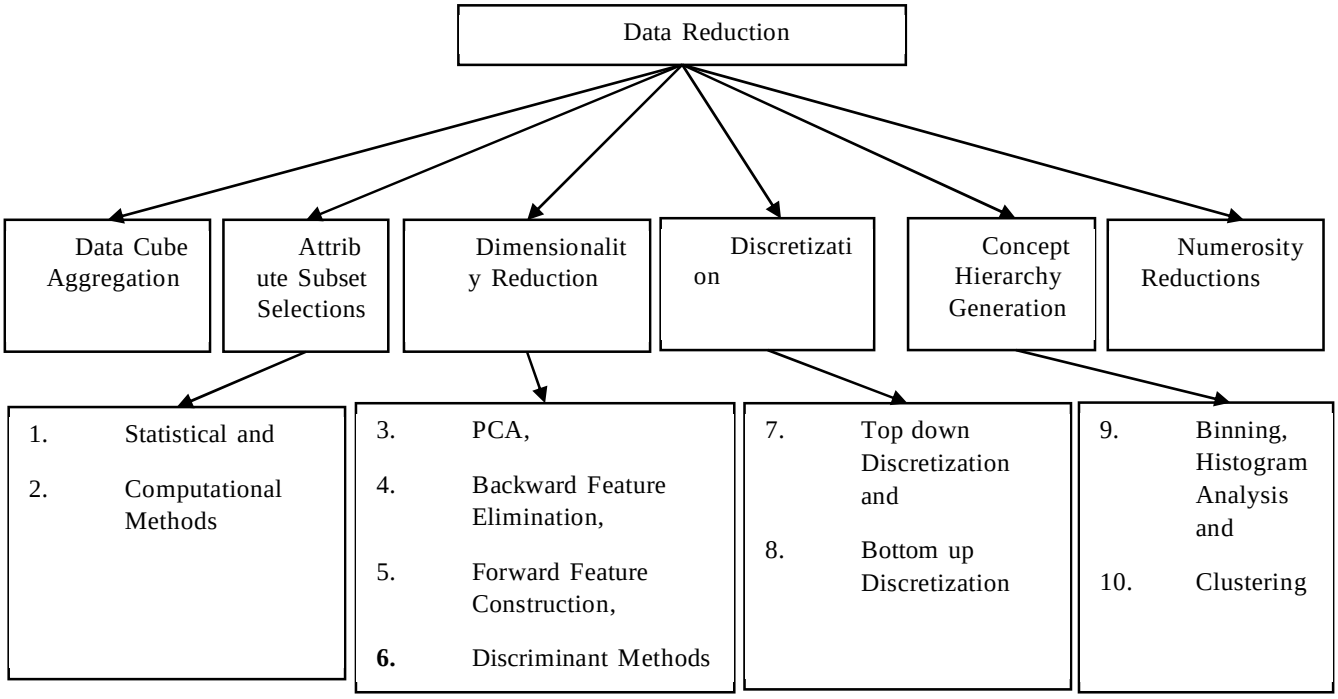


Figure: 4 Data Reduction Techniques

Data transformation is the process of converts' data from one format to another format. Data transformation includes various functions to achieve the perfect format that shown in fig 5.

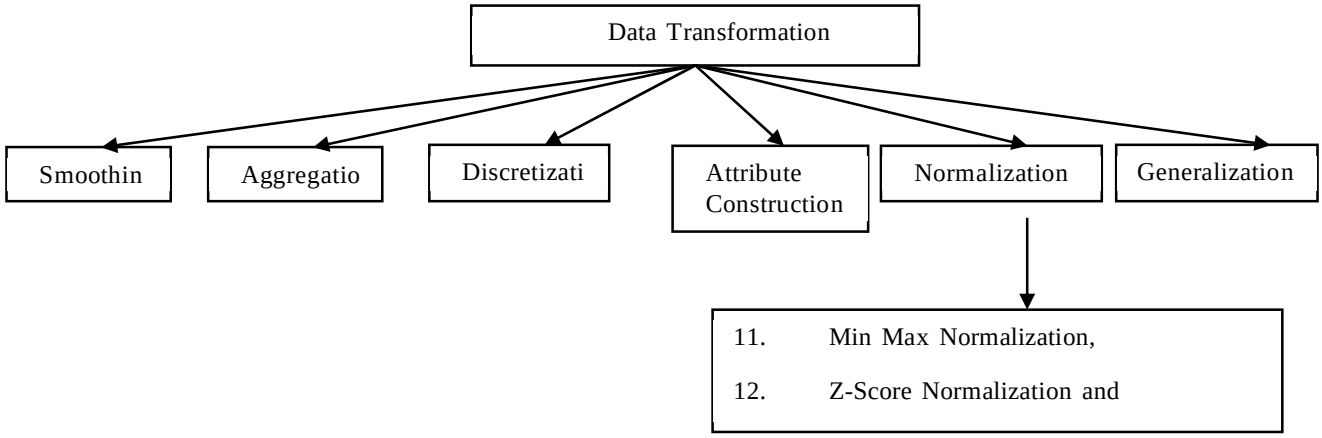


Figure: 5 Data Transformation Techniques

The above discussed techniques are mostly used for reducing the defects in dataset. By applying these techniques, the process of analytical models can be improved.

Moreover, the related work on preprocessing in various IoT-based applications are surveyed and listed in the following table 1.

Table 1: Uses of preprocessing Techniques in IoT-based Applications

Author Name & Year	Objective	Technique /Algorithm/Tool	Application Domain / Dataset
DiviyaPrabha 2016 [21]	Discuss various Technology that used in IoT for Data Collection and Data processing	Eclipse, KinomaJS, M2MLabs Mainspring, Node-RED, Raspberry Pi, RFID, QT (Quick Response), NFC (Near Field Communication), BLE (Bluetooth Low Energy), ZigBee	General IoT Environment

Peter 2017[22]	Overview Data Mining (DM) the Internet of Things (IoT), Preprocessing, Predictive Analytics	Machine Learning, Deep Learning, Natural Language Processing(NLP)	IoT Data
Brink2017 [23]	Provide solutions to Preprocessing problems	Modified Traditional Kalman Filter, intermittent Schmidt–Kalman filter (ISKF), the fixed-weight partial-update Schmidt–Kalman filter (FPSKF), and the partial-update Schmidt–Kalman filter (PSKF)	IMU camera Data
Bhavana 2017[24]	Survey & Discussion	IoT, Traditional Database management, Cloud, Sensor Data.	IoT Data
Shobanadevi 2017 [25]	Explain Role of Data Mining and Big Data in IoT	MapReduce, Appache Hadoop, KMeans, KNN(K nearest Neighbor), SVM(Support Vector Machine), Random Forest, Apriori	Health Care, Home Automation, Smart City
Akshat 2018[26]	Review	Data mining, IoT, Knowledge Discovery in Databases (KDD), Machine Learning	IoT Data
Pavithra 2019 [27]	Elaborate Role of Big Data in IoT to job and Market	Streaming Analytics, Spatial Analytics, Time Series Analytics, Prescriptive Analysis	General IoT Environment
Sandip 2019[28]	Survey	IoT, Radio Frequency Identification (RFID),Cloud, Machine to Machine Communication, Sensors and Actuators, Network Connectivity, Data Mining Preprocessing.	Smart application Data
Alcalde 2019 [29]	Library	Data Stream Library for Big Data Preprocessing DPASF	Streaming Big Data
Shivani 2019 [30]	Comparative Study	Reviewed all papers related with Noisy Data Between January 1993 to July 2018	Real world Data

Conclusion

Big data is now rapidly expanding across all domains such as education, agriculture, healthcare, institutions, web mining etc., Learning knowledge from this massive data is an interesting task as well as challenging one. Knowledge gaining from large sets of data brings significant opportunities and transformational potential to different sectors. But, the massive data comes with imperfection like noisy, missing values etc., this can lead to decrease in the efficiency and accuracy of decision making. So, refinement of data is required. This work offers the systematic flow of survey on data preprocessing techniques in the area of IoT and big environments. In which, the fundamentals of data preprocessing was covered, and literature reviews that related to the data preprocessing techniques were described. The classification of various pre-processing approaches with techniques was clearly depicted by the figure. Various approaches of preprocessing cleaning, transformation, reduction and integration with methods or techniques were illustrated. Data preprocessing techniques on various application were tabulated. Finally, issues and challenges, which need to be taken attention of in the future, were presented.

References

1. Bramer, Max. “Data for data mining”, In Principles of data mining*”*, pp. 9-19. *Springer*,*London*, 2016.
2. Alasadi, Suad A., and Wesam S. Bhaya.”Review of data preprocessing techniques indata mining, *Journal of Engineering and Applied Sciences* 12, no. 16 (2017): 4102-4107, 2017.
3. Cordón, Ignacio, Julián Luengo, Salvador García, Francisco Herrera, andFrancisco Charte. "Smartdata: Data preprocessing to achieve smart data in r." Neurocomputing 360, 1-13, 2019.
4. Hu, Hanqing, and Mehmed Kantardzic. "Smart preprocessing improves data streammining." In 2016 49th Hawaii International Conference on System Sciences (HICSS), pp.1749-1757. IEEE, 2016.
5. Shi, F.; Li, Q.; Zhu, T.; Ning, H.,“A survey of data semantization in internet of things”,*Sensors*, 18, 313,2018.
6. Shah, S. H., & Yaqoob, I, “A survey: Internet of Things (IOT) technologies, applications and challenges“,*IEEE Smart Energy Grid Engineering SEGE)*. doi:10.1109/sege.2016.7589556, 2016.
7. Teh, HuiYie, Kempa-Liehr, Andreas W, Wang, Kevin I-Kai, "Sensor data quality: a systematic review", *Journal of Big Data*, 7(1), 11–60,2020, doi:10.1186/s40537-020-0285-1
8. Weiss, Matthew, Wiederoder, Michael S, Paffenroth, Randy C, Nallon, Eric C, Bright, Collin J, Schnee, Vincent P, McGraw, Shannon; Polcha, Michael, Uzarski, Joshua R, "Applications of the Kalman Filter to Chemical Sensors for Downstream Machine Learning", *IEEE Sensors Journal*, (), 1–1, 2018, doi:10.1109/JSEN.2018.2836183
9. Hira, Z. M., &Gillies, D. F, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data", *Advances in Bioinformatics*, 1–13, 2015,doi:10.1155/2015/198363
10. Xu, C., Yang, H. H., Wang, X., &Quek, T. Q. S, "On Peak Age of Information in Data Preprocessing enabled IoT Networks", *IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, doi:10.1109/wcnc.2019.8885690
11. EvgeniyLatyshev, "Sensor Data Preprocessing, Feature Engineering andEquipment Remaining Lifetime Forecasting for PredictiveMaintenance", *Data Analytics and Management in DataIntensive Domains (DAMDID/RCDL’2018)*,226-231, 2018.
12. D. Natarajasivan and M. Govindarajan, "Filter Based Sensor Fusion for Activity Recognition using Smartphone", *International Journal of Computer Science and Telecommunications* Volume 7, Issue 5, 2016.
13. Morais, C. M. de, Sadok, D., &Kelner, J, "An IoT sensor and scenario survey for data researchers", *Journal of the Brazilian Computer Society*, 25(1), doi:10.1186/s13173-019-0085-7, 2019.
14. RajalakshmiKrishnamurthi, Adarsh Kumar, DhanalekshmiGopinathan, AnandNayyar, and Basit Qureshi, "An Overview of IoT Sensor Data Processing, Fusion, and Analysis Techniques", *Sensors*, 20, 6076; doi:10.3390/s20216076.
15. Gil, D., Johnsson, M., Mora, H., & Szymanski, J, "Advances in Architectures, Big Data, and Machine Learning Techniques for Complex Internet of Things Systems", *Complexity*, 1–3, doi:10.1155/2019/4184708,2019.
16. Gibert, Karina, Miquel Sànchez-Marrè, and Joaquín Izquierdo. "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining." *AI Communications* 29, no. 6 (2016): 627-663.
17. García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. "Big data preprocessing: methods and prospects." *Big Data Analytics* 1, no. 1 (2016): 9.
18. Hariharakrishnan, Jayaram, S. Mohanavalli, and KB Sundhara Kumar. "Survey of pre-processing techniques for mining big data." In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP),* pp. 1-5. IEEE, 2017.
19. Jamshed, Huma & Khan, M. & Khurram, Muhammad & Inayatullah, Syed & Athar, Sameen. (2019). Data Preprocessing: A preliminary step for web data mining. 206-221. 10.17993/3ctecno.2019.specialissue2.206-221.
20. Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey*, ”International Conference on Intelligent Sustainable Systems (ICISS)”,* ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.

21. V. Diviya Prabha R. Rathipriya, IoT Data and its Application-A Preliminary Study,"International Journal of Computational Intelligence and Informatics", Vol. 6: No. 1, 2016.
22. Peter Wlodarczak, Mustafa Ally, Jeffrey Soar,"Data Mining in IoT", *In Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17)*, ISBN 978-1-4503-4951, https://doi.org/10.1145/3106426.3115866, 2017.
23. Brink, K. M, " Partial-Update Schmidt–Kalman Filter", *Journal of Guidance, Control, and Dynamics*, 40(9), 2214–2228, doi:10.2514/1.g002808,2017.
24. Bhavana Bachhav, Parikshit N. Mahalle, "Data Management for Internet of Things: A Survey and Discussion", International Research Journal of Engineering and Technology (IRJET) ISSN: 2395-0056, Volume: 04, Issue: 11, 2017
25. Shobanadevi, A., & Maragatham, G. Data mining techniques for IoT and big data — A survey, *"International Conference on Intelligent Sustainable Systems (ICISS)",* ISBN:978-1-5386-1959-9,doi:10.1109/iss1.2017.8389260,2017.
26. Akshat Savaliya, Aakash Bhatia, Jitendra Bhatia, "Application of Data Mining Techniques in IoT: A Short Review", Volume 4, Issue 2, ISSN: 2395-1990,2018.
27. A.Pavithra, C.Anandhakumar, V.Nithin Meenashisundharam, Internet of Things with BIG DATA Analytics – A Survey, "International Journal of Scientific Research in Computer Science Applications and Management Studies IJSRCSAMS",Volume 8, Issue 1,ISSN 2319 – 1953, 2019.
28. Sandip Sonawane, "Survey on Technologies, uses and Challenges of IoT",International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 12, 2019.
29. Alcalde-Barros, A., García-Gil, D., García, S., & Herrera, F.,"DPASF: a flink library for streaming data preprocessing,"*Big Data Analytics*, 4(1). doi:10.1186/s41044-019-0041-8, 2019.
30. Gupta, S., & Gupta, A. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review." *Procedia Computer Science*, 161, 466–474. doi:10.1016/j.procs.2019.11.146, 2019.