1-1-2022

A Critical Study on the Effect of Dimensionality Reduction on Intrusion Detection in Water Storage Critical Infrastructure

Ranim Aljoudi
University of Windsor

A CRITICAL STUDY ON THE EFFECT OF DIMENSIONALITY REDUCTION ON INTRUSION DETECTION IN WATER STORAGE CRITICAL INFRASTRUCTURE

by

Ranim Aljoudi

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Master of Applied Science
at the University of Windsor

Windsor, Ontario, Canada

2021

A Critical Study on the Effect of Dimensionality Reduction on intrusion detection in Water Storage Critical Infrastructure

by

Ranim Aljoudi

APPROVED BY:

J. Ahamed,
Department of Mechanical, Automotive and Materials Engineering

---

R. Rashidzadeh,
Department of Electrical and Computer Engineering

M. Ahmadi, Co-Advisor
Department of Electrical and Computer Engineering

R. Razavi-Far, Co-Advisor
Department of Electrical and Computer Engineering

November 30, 2021

Declaration of Co-Authorship and Previous Publications

I. CO-AUTHORSHIP

I hereby declare that this thesis incorporates material that is the outcome of my research under the supervision of Dr. Majid Ahmadi and Dr. Roozbeh Razavi-Far. In general, the key ideas, methodology development, software programming, validation verification, data curation, writing, and visualization were performed by Ranim Aljoudi and the contribution of co-authors was primarily through the provision of study materials. Dr. Razavi-Far provided conceptualization ideas, conducting research, writing - review and editing, supervision, project management, and funding acquisition. Dr. Ahmadi assisted in investigation, conducting research, supervision, and funding acquisition. Dr. Mehrdad Saif helped in writing - review and editing. Mr. Ehsan Hallaji helped in writing - review and editing, formal analysis, and validation verification.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-authors to include the co-authored material in my thesis. I certify that, with the above qualification, this thesis and the research to which it refers is the product of my own work.

II. PREVIOUS PUBLICATION

This thesis partly includes the original papers that have been previously submitted, to be submitted, or published in peer reviewed journals and conferences as provided in the following table.

Thesis Chapter	Publication title/full citation	Publication status
Chapter 3	Aljoudi, R., Hallaji, E., Razavi-Far, R., Ahmadi, M., Saif, M., "A Study on the Effect of Dimensionality Reduction on Cyber-Attack Identification in Water Storage Tank SCADA Systems," in Explainable Artificial Intelligence within the context of Digital Transformation and Cyber Physical Systems, 2021, in press."	Published
Chapter 4	Hallaji, E., Aljoudi, R., Razavi-Far, R., Ahmadi, M., Saif, M., "A Critical Study on the Importance of Feature Selection for Diagnosing Cyber-Attacks in Water Critical Infrastructures," in Explainable Artificial Intelligence within the context of Digital Transformation and Cyber Physical Systems, 2021, in press."	Published

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor. I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis. I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Supervisory control and data acquisition (SCADA) systems are often imperiled by cyber-attacks, which can often be detected using intrusion detection system (IDSs). However, the performance and efficiency of IDSs can be affected by several factors, including the quality of data, curse of dimensionality of the data, and computational cost. Feature reduction techniques can overcome most of these challenges by eliminating the redundant and non-informative features, thereby increasing the detection accuracy. This study aims to shows the importance of feature reduction on the intrusion detection performance. To do this, a multi-modular IDS is designed that is connected to the SCADA system of a water storage tank. A comparative study is also performed by employing advanced feature selection and dimensionality reduction techniques. The utilized feature reduction techniques improves the IDS efficiency by reducing the memory usage and using data with better quality, which in turn increase the detection accuracy. The obtained results have been analyzed in terms of F1-score and accuracy.

Acknowledgements

This work would not have been possible without the help and support of many people. Writing this thesis has had a big impact on me. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level. I would like to reflect on the people who have supported and helped me so much throughout this period.

First and foremost, I would like to express my sincere gratitude to my co-supervisor, Dr. Roozbeh Razavi-Far, for his guidance and unwavering support throughout my research. He was always spot willing to answer any questions or concerns I had and provided with the opportunity to be creative in my field of study. In addition, express my gratitude to my co-supervisor Dr. Majid Ahmadi, Associate Dean of Graduate Studies at Faculty of Engineering. I thank him heartily for his encouragement, massive academic support and granting me so many wonderful opportunities.

My appreciation is further extended to my thesis committee members, Dr. Rashid Rashidzadeh from the Electrical and Computer Engineering Department, and I am also hugely appreciative of Dr. Jalal Ahmed of Mechanical, Automotive and Materials Engineering for agreeing to be an external reader and his invaluable input and time. Their critical evaluation improved this work to a large extent. Without their passionate participation and input, this thesis could not have been successfully conducted.

I am also grateful to Ms. Andria Ballo, graduate secretary at the Electrical and Computer Engineering Department, for her time and support.

Nobody has been more important to me in the pursuit of this thesis than the members of my family. I would like to thank my whole family member, whose love and guidance are with me in whatever I pursue. They taught me how to find happiness in even the most difficult situations; they are my role models.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

SCADA	Supervisory Control And Data Acquisition
IDS	Intrusion Detection System
IOC	Indicators Of Compromise
SIDS	Signature Intrusion Detection System
AIDS	Anomaly Intrusion Detection System
NIDS	Network Intrusion Detection System
HIDS	Host Intrusion Detection System
PIDS	Protocol-based Intrusion Detection System
APIDS	Application Protocol-based Intrusion Detection System
HIDS	Hybrid Intrusion Detection System
DT	Decision Tree
KNN	K-Nearest Neighbor
FE	Feature Extraction
FS	Feature Selection
FR	Feature Reduction
DR	Dimensionality Reduction
PCA	Principle Component Analysis
FA	Factor Analysis

CFA	Confirmatory Factor Analysis
MDS	Multidimensional Scaling
LDA	Linear Discriminant Analysis
ISO	Isomap
SM	Semantic Mapping
PPCA	Probabilistic Principal Component Analysis
LLE	Locally Linear Embedding
LE	Laplacian Eigenmaps
LIM	Landmark Isomap
HLLE	Hessian-based Locally Linear Embedding
LTSA	Local Tangent Space Alignment
KPCA	Kernel PCA
GDA	Generalized Discriminant Analysis
NPE	Neighborhood Preserving Embedding
LPP	Locality Preserving Projections
DM	Diffusion Maps
LLC	Locally Linear Coordination
MC	Manifold Charting
LMNN	Large Margin Nearest Neighbour
ICA	Independent Component Analysis

InfFS The infinite feature selection

ILFS Infinite latent feature selection

ECFS Evolutionary computation feature selection

ReliefFS Relief Feature Selection

MutlnfFS Mutual information

mRMR Maximum Relevance and Minimum Redundancy

FSV Feature Selection via Concave Minimization

Laplacian Laplacian score

MCFS Multi-Cluster Feature Selection

RFE Recursive Feature Elimination

Chapter 1

Introduction

1.1 Background

Recent advancements in cyber-physical systems are often followed by more dependency on the application layer [1, 2, 3]. The severity of the intrusions to computer networks has been gradually increasing by threatening the security of these networks through violating privacy, integrity, and accessibility mechanisms [4, 5].

Plenty of features in cyber-physical systems can complicate explaining various events for various applications. Security challenges arise from two different perspectives. Firstly, an event may be detectable, when a change happens in a certain combination of features, and as the number of these variables and events increases, explaining the system status becomes more difficult. Furthermore, the feature space may contain hidden characteristics that cannot be seen by the naked eye. Devising dimensionality reduction (DR) technique improves the clarity of such systems in a number of ways[6, 7, 8, 9]. These techniques aim at improving the feature space by capturing the complex structure of the original data, and, then, transform it into a low-dimensional space, which facilitates visualization, thereby revealing relationships between samples, understanding and monitoring the dynamics of the system.

Intrusion detection system usually rely on prior knowledge, training data, or recorded data, which is often complex to analyze for extracting the attack pattern. When dealing with big data, the high dimensionality of the data, which is the focus of this work, complicate the decision-making process, as it severely decrease the efficiency of the system and quality of the constructed model. Moreover, industrial

datasets usually contain noisy, redundant, or irrelevant features that introduce critical challenges to data modeling [10]. Feature Selection (FS) techniques can be used to tackle the high dimensionality of the data and address the low quality of the data by removing redundant and non-informative features [11, 12]. Such improvement in data quality will in turn enhances the performance of data-driven modules such as change detectors [13, 14] and classifiers [15, 16, 17, 9] in the system.

1.2 Intrusion Detection

An Intrusion Detection System (IDS) is a security mechanism to inspect traffic via detecting and tackling computer security threats or any suspicious behavior [4, 5]. IDS monitors the system/network, and, then, detects intrusions and the occurrence of cyber-attacks . It has been widely used in recent years as one of the main network security components and cyber-physical systems [18, 3, 19]. Various challenges are arising in accurately detecting intrusions, which make the majority of studies cyber-physical systems to focus on more advanced approaches such as machine learning [19, 3]. An intrusion detection system can also be characterized as a device or an application that detects malicious activities within the network. To secure industrial network systems [1, 2, 3], we require to address malicious intrusions that are violating privacy, integrity, and accessibility. A major challenge in current IDSs is the high dimensionality of the network data so that the classifiers cannot distinguish the normal behavior of the system accurately and in a timely manner due to the existence of irrelevant and redundant features. Therefore, IDSs experience lack of prediction, high computational overheads and tardy detection. Furthermore, due to the massive number of feature subsets that can be selected from input features, depending on the feature set dimensionality, it is challenging, if not impossible to use a comprehensive search and test each and every subset[10].

SCADA systems are used for monitoring and controlling various critical infrastruc-

ture processes through receiving data from sensors [20, 18]. It controls the mechanical machines, while the software allows human interactions to manage the machines. A traditional IDS needs a database that holds records of different attacks, in which each record corresponds to a particular intrusion and its characteristics. The major drawback of this mechanism is the necessity for human involvement to inspect threats, which is a very complicated and time consuming task. Therefore, it is necessary to use machine learning techniques to promote anomaly detection algorithms that can discover abnormal changes in the system [13, 18].

1.2.1 Model-based Approaches

The main benefit of model-based approaches is discovering anonymous attacks. IDS is generally classified into five types: Network Intrusion Detection Systems (NIDS), Host Intrusion Detection Systems (HIDS), Protocol-based Intrusion Detection Systems (PIDS), Application Protocol-based Intrusion Detection Systems (APIDS), and Hybrid Intrusion Detection Systems. NIDS are set up at a planned point within the network to examine traffic from all devices on the network. NIDS can check several hosts simultaneously, and it is capable of detecting the broadest ranges of network protocols. HIDS run on independent hosts or devices on the network, in which they monitors the incoming and outgoing packets from the device only. Intrusions are detected by HIDS through checking hosts file system, system calls or network events. Protocol-based intrusion detection systems (PIDS) consist of a system or agent that would consistently reside at the front end of a server, controlling and interpreting the protocol between a user/device and the server. PIDS aim to secure the web server by regularly monitoring the HTTPS protocol stream and accept the related HTTP protocol. APIDS are those agents that settle within a group of servers and identify the intrusions by monitoring and interpreting the communication on application specific protocols. APIDS should be placed between a web server and the database management system in order to monitor a particular SQL protocol to the business logic

because it interacts with the database. Lastly, HIDS are made by the combination of two or more approaches of the intrusion detection system [21].

1.2.2 Data-driven Approaches

IDS can also be classified based on the input data sources used to detect abnormal activities. The most common approaches are signature-based (SIDS) and anomaly-based approaches (AIDS). The signature-based approaches; also known as Knowledge-based Detection or Misuse Detection, refer to the detection of attacks by looking for specific sequences or patterns, such as byte sequences in network traffic, that match a particular attack signature. Also, they are known as effective approaches for detecting known attacks because they rely on a prerecorded list of known indicators of compromise (IOCs) and they capture the recognized properties of the attacks. Alternatively, when an intrusion signature matches with the signature of a prior intrusion that already exists in the signature records, a triggered alarm is activated. The major limitation of the signature-based approaches is that they can only detect the intrusions whose attack patterns are already stored in the database. Therefore, the efficiency of SIDS decrease due to updating dynamical signatures for various patterns.

Anomaly-based intrusion detection systems (AIDS) have overcome the limitation of SIDS and are being used to identify malicious attacks on systems. AIDS monitor network traffic and compare it with a predefined baseline that is considered normal for the network concerning bandwidth, protocols, and ports. AIDS use machine learning techniques to establish a model of the normal behavior and accompanying security policy. Therefore, any significant deviation between the observed behaviour and model behaviour is classified as anonymous intrusions. The main advantage of AIDS is the ability to identify zero-day attacks due to the fact that recognizing the abnormal user activity does not rely on a signature database [22]. Generally, AIDS could be used to detect new attacks and create intrusion signatures.

1.3 Intelligent methods for detecting and classifying cyber-attacks

Many researchers used machine-learning techniques for classifying cyber attacks. Computational security models are developed to inspect several cyber event patterns and at the end they anticipate the threats employing cyber security data that can be applied for building a data-driven intelligent IDS. Machine learning techniques can be applied for constructing such a data-driven intelligent approach for the intrusion detection system. Popular machine learning classification algorithms, like Bayesian Networks, Naive Bayes classifier, Decision Tree, Random Decision Forest, Random Tree, Decision Table, and Artificial Neural Networks, are among those methods that are implemented to detect intrusions to provide intelligent services in the domain of cyber-security. Such as Amiri et al. [23] used a least-squared support vector machine classifier to train the model utilizing large datasets to create a speedy and efficient predictive model for classifying intrusions. A probability-based Bayesian network was used to classify the events that process TCP/IP packets [24]. Koc et al. [25] build a naive Bayes classifier to build a multi-class intrusion detection system. Classification algorithm is the core element of any intrusion detection, the selection of right classifier plays an important role in the detection accuracy and the overall performance of the intrusion detection algorithm. The KNN classier is the most popular machine learning method, in which the classification of a point is determined by the k-nearest neighbours of another data point. Researchers like Shapoorifard et al. [26] and Vishwakarma et al. [27] used the KNN classification technique in their studies for intrusion detection. Moreover, the decision tree classification approach is significantly used in experiments for building intrusion detection systems. A logistic regression model was used for identifying malicious traffic and intrusions by sampling the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Lastly, various performance metrics including precision, recall, F1-

score, and accuracy,have been used to evaluate the effectiveness of predictive models used for clarifying cyber-attacks and intrusions.

1.4 Impact of feature reduction on intrusion detection

High-dimensional datasets lead to negative impacts on the performance and computational time of machine learning-based IDSs. To address this issue, numerous researchers applied data processing techniques such as feature reduction. Therefore, feature reduction is an important step for detecting intrusions and classifying cyber-attacks. Reducing the number of relevant traffic features without a negative impact on the classification accuracy is a target that extremely benefits the overall effectiveness of an intrusion detection system. Feature reduction can be used in selecting relevant features for building durable IDSs models and can be effective on both efficiency and performance of the IDS models. To look into intrusive patterns in the IDS database, some of the features in the network are redundant and irrelevant. This increases the processing time and lowering the performance of IDS; therefore, we require to remove the useless features from the original high dimensional database.

1.4.1 Dimensionality reduction and feature selection

The quality of data in a data-driven process is usually affected by various factors [28, 29]. One of the most common challenges in machine learning is the issue of high-dimensionality, which can be addressed by feature reduction. Feature reduction is mostly used for data analysis, compression, and data visualization. Most of the feature reduction methods are divided into two main categories: (i) Feature selection: approaches select a subset of features from the original feature space that results in the optimal performance. (ii) In contrast, dimensionality reduction, also called

feature extraction, captures the structure of the original feature space, and, then, transforms into a lower dimensional features space. Dimensionality reduction is the process of improving the original feature space and transforming it into a smaller one to minimize the complexity of a model and avoid the curse of dimensionality [16, 30]. This data transformation may be linear or non-linear.

Feature selection, also known as variable selection or attribute selection, is a common technique for improving data quality. This process obtains a subset of relevant features and eliminates the irrelevant and redundant features from the original data. The main difference between feature selection and dimensionality reduction is that the former creates space by adopting a subset of features from the original feature space, while the latter transforms the original feature space and creates a completely new feature space. Feature selection can improve the accuracy of the model, reduce learning time, and prevent overfitting. Most feature selection methods are divided into three major buckets: (i) filter-based: generally, analyzes intrinsic properties of data, regardless of the classifier. It only considers the association between the feature and the class labels. (ii) wrapper-based: this method is based on a specific machine learning algorithm to find optimal features; it uses classifiers to score a given subset of features. (iii) embedded: is an iterative method, in which the selection process is employed for the learning of the classifier. Most of these methods can perform two processes: ranking and subset selection (sometimes they are performed sequentially): the importance of each feature is evaluated, usually by neglecting potential interactions among the elements of the joint set, then the final subset of features to be selected is provided.

While feature selection techniques often operate singularly and are not combined with other feature selection algorithms, it is also possible to use these techniques in combination. By doing so, one can use simple approaches such as the majority of votes to aggregate the results of these techniques. However, this approach will be most advantageous when the selected algorithms employ completely different methods

(e.g., manifold learning, cluster analysis, and mutual information) to capture the distribution of the feature space. This will extend the flexibility of the feature selection process against various distribution types and data structures.

Dimensionality reduction can be very helpful in the design of intrusion detection systems (IDS). For instance, if a cyber-attack can be detected by monitoring a large number of features, dimensionality reduction can yield a feature space, in which only one or a very small number of features are enough to explain a change that indicates a cyber-attack. In contrast, other techniques such as feature selection may not result in the same efficiency, as the features may not have enough information to only select a small number of them to detect a cyber-threat. In other words, feature selection usually works when at least several features possess very useful information, while dimensionality reduction tries to rectify the feature space and obtain an improved feature space.

1.5 Outline

The subsequent chapters of this study are structured as follows:

Chapter 2 discusses Water Storage System by explaining it's main components, characteristics of collected data including classes attack types, and the design of the intrusion detection system. In **Chapter 3**, we will explore the effect of dimensionality reduction (DR) on the classification accuracy of cyber-attacks in the cyber-physical system. We will conduct our experiments with 22 advanced feature extraction models combined with two standard classifiers, K-Nearest Neighbours (KNN) and Decision Tree (DT), which are expected to effectively select the optimal set of features for classifying cyber-attacks. **Chapter 4** explains the effect of Feature Selection techniques for classifying cyber-attacks in water critical infrastructures. Lastly, **Chapter 5** compares the results attained through feature selection and diensionality reduction and shows, which methods perform best in our water storage system.

Chapter 2

Problem Statement

2.1 Water Storage System

The SCADA datasets have been used for the evaluation of the intrusion detection. In this case study, SCADA was implemented for a water storage tank system. An intruder can hack into the network system of this cyber-physical system and disrupt the operation of the control unit. SCADA systems are generally made of four groups of components. The first components are the sensors and actuators that collect data from remote facilities. These data have information about the state of the physical process. By this means, commands can be sent to control the physical process and create a feedback control loop. Secondly, the programmable logic controllers are pointed to remote terminal units (RTUs) to collect data, which define the system's state. The water tank RTU ladder logic includes six setpoint registers; HH (High-High), HI (HIgh-Low), LO (Low), and LL (Low-Low) water level setpoint register, a pump override setpoint register, and a mode setpoint register. Furthermore, it includes three output registers, which store process parameters; pump state, water level, and alarm state. Thirdly, supervisory controls are handled by the master terminal unit (MTU), which in turn forwards commands to RTU. MTU sends a read query to read from the registers to measure the state of the system. The fourth level is the human-machine interface (HMI) that is used to display the sensor data received by MTU. HMI provides an interface for an operator to monitor and control the system and operations in the form of visual representation.

HMI supports a communication protocol such as MODBUS commands. HMI

(master) sends commands to MODBUS (slave), in which the individual RTU executes the command and returns a response. MTU copies commands and responses received from the HMI port to the radio port and vice versa, whereas the HMI software makes changes (every 2 seconds) to setpoint register values to control the physical process [31].

2.2 Data Characteristics

The intrusion detection system detects and classifies seven different types of cyber-attacks in the water storage tank system, along with the normal class when the system is safe and an injected class, as shown in Table 2.1. The network traffic data are used for training and validation of predictive models that are integration of state-of-the-art DR and FS techniques to construct a signature-based intrusion detection systems. After the traffic data is passed through dimension reduction and feature selection techniques, the most relevant features are selected, and new reduced set of features is used for training classifiers. These predictive models classify seven different types of cyber-attacks in the Water Storage System along with the normal class. These classes are reported in 2.1. These injection attacks are also explained

Table 2.1 – List of classes including normal and simulated cyber attacks in Water Storage System.

Classes	Class type: Normal status and attacks
0	Instance not part of an injection
1	Naive Malicious Response Injection
2	Complex Malicious Response Injection
3	Malicious State Command Injection
4	Malicious Parameter Command Injection
5	Malicious Function Code Injection
6	Denial Of Service Injection
7	Reconnaissance Injection

briefly in the following:

- Naïve Malicious Response: can be used to send fake payloads by injecting response packets into the network.
- Complex Malicious Response: conceals the state of the controlled physical process to maliciously affect the feedback control loop.
- Malicious State Command: manipulates remote field devices to change the normal system state to a critical state by sending malicious commands.
- Malicious Parameter Command: mainly tries to change the set-points defined for programmable logic controllers.
- Malicious Function Code: refers to the commands included in the application layer of a system, which can be used maliciously by attackers to create unintended consequences.
- Denial of Service: corrupts communications links and system programs by attempting to exhaust computational resources.
- Reconnaissance: is the process in which attackers gain device information and system vulnerabilities to plan future attacks against a SCADA system.

The network traffic data is recorded from MODBUS traffic with an RS-232 connection, in which it is one byte long and each server has a unique device address. The water system contains a relief valve to drain water from the tank, a pump, alarm, meter, and a switch control scheme to maintain the water level between high and low setpoints.

An attack can be observed by the read and write commands/responses, which have a fixed length for each system. In a normal system, the error rate should be low and constant but when the system undergoes a denial-of-service attack the rates are expected to increase. If there is no error during the normal state, the response function code matches the command function. When there exists an error, the response sub-

function code equals the summation of the command function code and a value of 0X80.

Table 2.2 – List of raw features in the Water Storage System.

Number	Feature Name	Description
1	Command address	Device ID in command packet.
2	Response address	Device ID in response packet.
3	Command memory	Memory start position in command packet.
4	Response memory	Memory start position in response packet.
5	Command memory count	Number of memory bytes for R/W command.
6	Response memory count	Number of memory bytes for R/W response.
7	Command read function	Value of read command function code.
8	Command write function	Value of write command function code.
9	Response read function	Value of read response function code.
10	Response write function	Value of write response function code.
11	Sub-function	Value of sub-function code in the command/response.
12	Command length	Total length of command packet.
13	Response length	Total length of response packet.
14	H	Value of H set-point.
15	HH	Value of HH set-point.
16	L	Value of L set-point.
17	LL	Value of LL set-point.
18	Control mode	Automatic,manual or shutdown.
19	Control scheme	Control scheme of the water pipeline.
20	Pump	Value of pump state.
21	CRC rate	CRC error rate.
22	Measurement	Water level.
23	Time	Time interval between two packets.
24	Result	Manual classification of the instance.

The water storage tank system generates network flow records that are captured with a serial port data logger that includes 200,000 samples recorded using a laboratory-scale test-bed. 19503 of these samples correspond to the normal state

(i.e., class 0), and the rest of the samples are collected when the system was under attack. Classes 1 to 7 in Table 2.1 have 1198, 1457, 209, 410, 155, 135, 4132 samples, respectively. To detect malicious activities in the Water Storage System, features were divided into network traffic features and payload content features. The former gives information regarding the communications within the SCADA network system, while the latter describes the current state for different components of the SCADA system. The developed dataset consists of 24 unique features (i.e., 8 payload and 16 network traffic features) as shown in Table 2.2.

2.3 Design of the Intrusion detection system

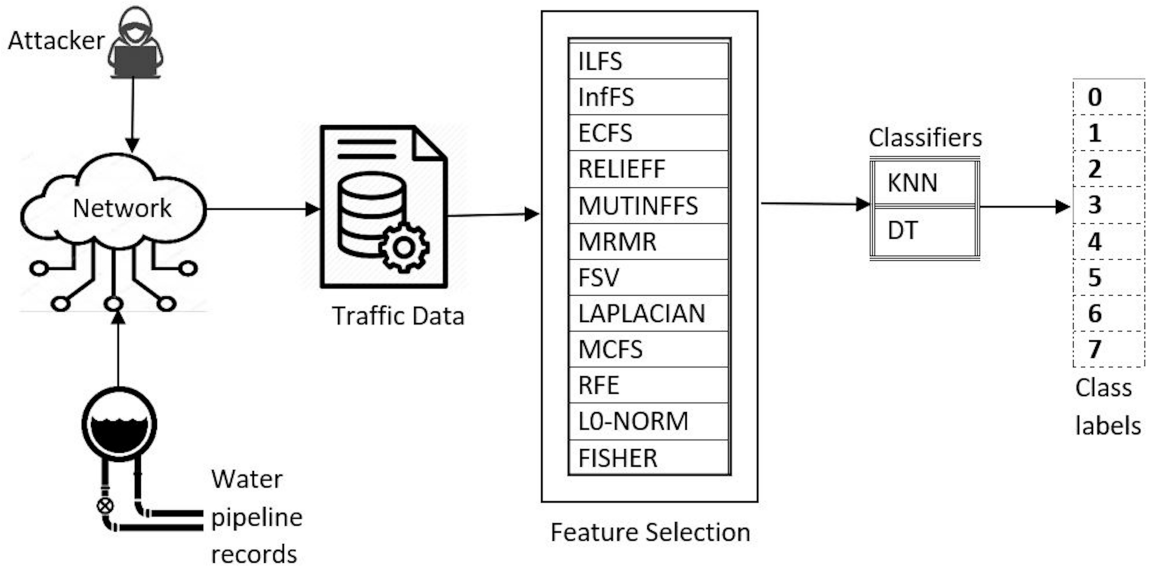

Figure 2.1 – Illustrative diagram of the designed feature selection-based system.

The designed Intrusion detection system (IDS) uses a multi-modular structure, in which the traffic data initially passes through FS and DR methods. Then, the selected features of data will be passed to the classification module, where the normal and malicious traffic can be classified based on their type (see Fig.2.1 for FS-based

scheme). Fig 2.2 refers to the dimensionality reduction-based scheme for classifying cyber-attacks.

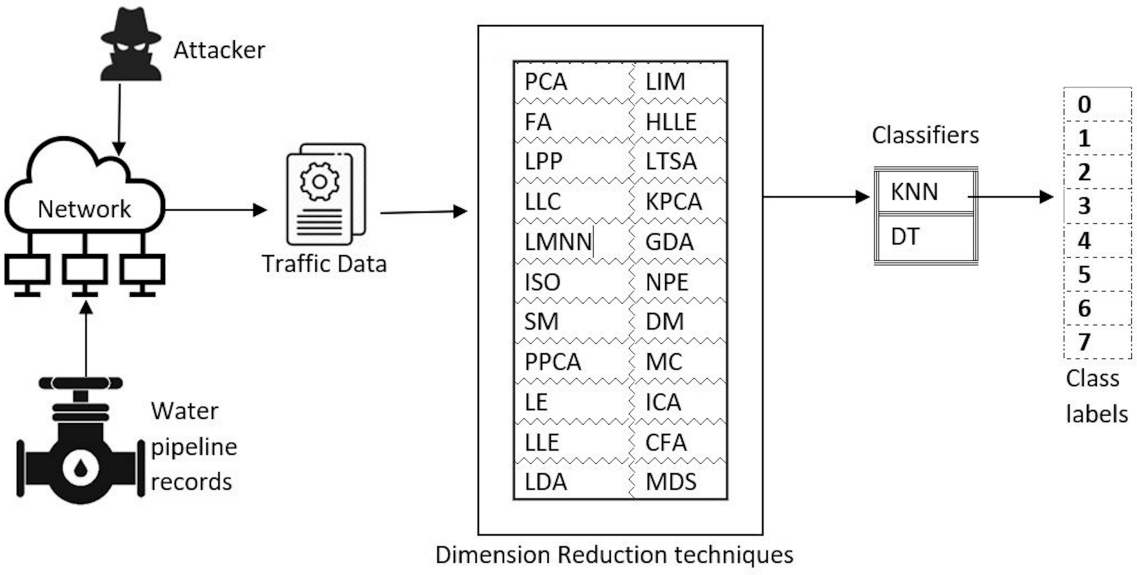

Figure 2.2 – Illustrative diagram of the designed dimensionality reduction-based system

2.3.1 Data Collection

The data is collected from this cyber-physical system that resembles a water storage tank [32]. SCADA systems collect data from remote facilities about the state of the physical process and send commands to control the physical process creating a feedback control loop. SCADA was used to control this water storage tank as it has communication patterns that are set of repetitive read and write commands. First, it writes the contents of all registers that are used for the control. Then, a MODBUS protocol reads the holding register command that measures the state of the system. This protocol acts as a single serial cable that connects the serial ports on Master and Slave devices. These two commands are each followed by a response. The raw collected data consisted of variables such as command and response address,

command and response memory, command and response memory count, command read and write function, the response read and write function, sub-function, command length, response length, control mode, control scheme, high (H) set-point value, high alert (HH) set-point value, low (L) set-point value, low alert (LL) set-point value, pump state, cyclic redundancy code (CRC) error rate, water level measurement, time-stamps, and the attack class.

2.3.2 Decision-Making

In our case study, the decision tree (DT) and K-nearest neighbours (KNN) algorithms are employed as classifiers in combination with multiple dimensionality reduction and feature selection techniques. DT is a classification method that uses a representation of a tree structure consisting of internal nodes that represent a test on an attribute and branches, which denote the outcome of the test and each leaf node holds a class label. KNN is a supervised learning algorithm (non-parametric algorithm) that uses the label information to learn new unlabeled data based on a similarity measure by calculating the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance. DT and KNN classifiers can analyze data and identify significant characteristics in the network.

Chapter 3

Effect of Dimensionality Reduction on Intrusion Detection

Industrial data-driven models are often challenged with various obstacles [33, 34]. One of the most common challenges in machine learning is the issue of high-dimensionality, which can be addressed by dimensionality reduction. Dimensionality reduction is the process of improving the original feature space and transforming it into a smaller one in order to minimize the complexity of a model and avoid curse of the dimensionality [16, 30]. Dimensionality reduction is mostly used for data analysis, compression, and data visualization.

3.1 Review of Dimensionality Reduction Methods

Dimensionality reduction, capture the structure of the original feature space, and, then, transform onto a lower dimensional features space. This data transformation may be linear or non-linear. The focus of this chapter is on the dimensionality reduction, and the selected techniques are explained in the following subsections. In contrast to feature selection techniques that may be used in combination to provide different rankings for the features, dimensionality reduction techniques are preferred to be used alone. This is due the fact that the created features spaces may represent different distributions and do not share any common features. A question, however, may arise regarding the criteria for selecting the right technique for the task at hand. While various measures can be used to facilitate this decision, the best choice could be made after testing different algorithms on the same data to see which one is more

adaptable with the case study and can result in a higher performance. This is the approach followed in this chapter. Nevertheless, should one desire to choose a versatile technique that works with various case studies, there are a few points to consider. Firstly, it is more desirable to use supervised dimensionality reduction methods, if labeled data is available, as their valuable information will be discarded by unsupervised methods. Secondly, many of dimensionality reduction methods make use of approaches such as manifold learning and kernel functions. These techniques are very powerful, if they are carefully optimized, and the data distribution should match the underlying assumptions such as comparability with the selected kernels or to be projectable onto a manifold.

3.1.1 Principal Component Analysis

Principle Component Analysis (PCA) is a very established method, as an unsupervised linear transformation technique. PCA supports us to identify patterns in the data based on the correlation between features. PCA projects the direction of maximum variance in high-dimensional data onto a lower-dimensional subspace in order to minimize the sum of squared error, or maximizes the variance. It is decomposed by obtaining eigenvectors and eigenvalues on the data covariance matrix of the whole dataset. The obtained eigenvalues represent the variance of the projected inputs along principal axes, and eigenvectors (principal components) determine the directions of the new feature space. The benefits of PCA include the reduction of noise in the data and the ability to produce independent and uncorrelated features [35].

3.1.2 Factor Analysis

Factor Analysis (FA) is a statistical method that can be considered as an extension of PCA. FA is designed to identify the unobservable variables from the observed patterns of correlation between the variables. This is in contrast to PCA, as it is unable to

use the observed information. A factor is correlated with multiple observed variables, so each factor describes an appropriate amount of variance in the observed variables [36].

3.1.3 Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is a multivariate statistical method that measures variables representing the number of constructs (or factors). CFA models the data density on a low-dimensional manifold on which the data is representable [37]. CFA also follows a global approach for parameter optimization of the manifold estimation, which results in a satisfying convergence rate.

3.1.4 Multidimensional Scaling

Multidimensional Scaling (MDS) references the overall similarity (or dissimilarity) of the objects. MDS is used to visualize the dissimilarities or distances (usually by Euclidean distance) between objects by projecting the points to a low dimension space [38].

3.1.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised linear transformation that reduces the dimensionality on multi-class data by linearly projecting the original samples to a smaller space, while maintaining the class-discriminatory characteristics of the original data [39].

3.1.6 Isomap

Isomap (ISO) is also referred to as isometric mapping, it is a non-linear dimensionality reduction method, which takes the advantage of local information by using the concept

of geodesic distances induced by a neighborhood graph. This graph is embedded between pairs of points rather than Euclidean distances [40].

3.1.7 Semantic Mapping

Semantic Mapping (SM) reduces the dimensionality by clustering the original co-occurrent features. Using these semantic clusters and combining features mapped in the same cluster, it then generates an extracted feature that contains semantically related terms [41].

3.1.8 Probabilistic Principal Component Analysis

Probabilistic Principal Component Analysis (PPCA) offers an extension to the scope of PCA. PPCA can be utilized as a Gaussian model by maximizing the likelihood estimates of the parameters that are associated with the covariance matrix and can be efficiently computed from the data principle component [42].

3.1.9 Locally Linear Embedding

Locally Linear Embedding (LLE) is an unsupervised learning algorithm and a non-linear dimensionality reduction technique. LLE outlines its inputs into a single global coordinate system of lower dimensionality without the involvement of local minima. By employing the local symmetries of linear reconstructions, it can study the global structure of non-linear manifolds. LLE projects the points to a locally linear neighborhood. LLE utilizes an eigenvector based optimization technique to find the low-dimensional embedding of points [35].

3.1.10 Laplacian Eigenmaps

Laplacian Eigenmaps (LE) is a closely related approach to LLE. LE constructs a graph to compute a low-dimensional representation of the dataset that preserves local

neighborhood constraints of the dataset in an optimal process. LE is constructed by a weighted graph with k nodes. Each data point is a node, and a set of edges connecting the proximity of neighboring points using the K-nearest neighbor algorithm [43].

3.1.11 Landmark Isomap

Landmark Isomap (LIM) is a variant of Isomap that selects a group of points termed as landmarks to simplify the embedding computation. LIM only computes the shortest path from each data point to the landmark points. The classical MDS is then applied to the resulting geodesic distance matrix to find a Euclidean low-dimensional embedding of all data points [44].

3.1.12 Hessian-based Locally Linear Embedding

Hessian-based Locally Linear Embedding (HLLE) may be considered as an improved version of the LLE. Its theoretical approach is somehow similar to the Laplacian eigenmap framework, if the Laplacian operator is replaced with the Hessian. HLLE uses orthogonal coordinates on the tangent planes. This makes the local fits more robust for the dimensionality reduction [45].

3.1.13 Local Tangent Space Alignment

Local Tangent Space Alignment (LTSA) uses manifold learning, which can convert a non-linear embedding of high dimensional data into a smaller space, and rebuild high-dimensional coordinates from embedding coordinates. The steps for performing LTSA are similar to LLE; however, it is different in optimizing the embedding. In LTSA, we compute the tangent space of each data point and align those local tangent spaces, while ignoring the label information [46].

3.1.14 Kernel Principal Component Analysis

Kernel PCA (KPCA) is an extension of PCA for performing non-linear dimensionality reduction through the use of kernels. PCA can be applied to datasets that are linearly separable. This is while kernel PCA maps non-linear datasets and uses a kernel function (also called non-linear mapping function) to project dataset onto a higher dimensional feature space, where it is linearly separable [47].

3.1.15 Generalized Discriminant Analysis

Generalized Discriminant Analysis (GDA) is designed for a non-linear transformation. It utilizes kernel functions to map the data onto a new space, which leads to non-linear discriminant analysis for the input data. This has been done by maximizing the ratio of the between-class scatter to the within-class scatter [48].

3.1.16 Neighborhood Preserving Embedding

Neighborhood Preserving Embedding (NPE) is a linear DR method that aims to discover the local neighborhood structure on the data manifold. Each data point is represented as a linear combination of the neighboring data points and coefficients that are specified in the weight matrix. It then finds an optimal embedding such that the neighborhood structure can be preserved in the resulted feature space [49].

3.1.17 Locality Preserving Projections

Locality Preserving Projections (LPP) is similar to NPE in aiming at preserving the local manifold structure. LPP shares a lot of LE or LLE properties. LPP employs the concept of non-linear Laplacian eigenmap and computes a transformation matrix that maps the data points to a new space. The projective maps in LPP are the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold [50].

3.1.18 Diffusion Maps

Diffusion Maps (DM) reduces the data dimensionality by re-arranging data according to parameters of its underlying manifold. The Euclidean distance between points in the embedded space is equal to the diffusion distance in the original dimension space. The connectivity between the points is measured using a local similarity measure at different scales [51].

3.1.19 Locally Linear Coordination

Locally Linear Coordination (LLC) computes a number of locally linear models on data using the Expectation Maximization approach. By this mean, it performs a global alignment of the linear models by aligning the local linear models using a LLE variant [52].

3.1.20 Manifold Charting

Manifold Charting (MC) minimizes a cost function that measures the amount of difference between the linear models on the global coordinates of the data points by solving a generalized eigenproblem [52]. MC also shares some similarities with the LLC technique.

3.1.21 Large Margin Nearest Neighbour

Large Margin Nearest Neighbour (LMNN) is based on semi-definite programming for optimizing a convex problem. The target neighbors can be set as a k-nearest neighbors rule that shares the same labeled instances. The new data instances are obtained from the highest vote of the k closest labeled instances. Using the global distance metric learning method, it measures the nearby instances from the same class and eliminates instances from different classes [53].

3.1.22 Independent Component Analysis

Independent Component Analysis (ICA) is a computational method that transforms the independent components of the observed data by increasing the statistical independence of the estimated components. ICA aims to separate multivariate signals into components that are maximally independent of each other by applying a linear transformation to decompose the original data. ICA aims to increase the accuracy for uncorrelated data; however, the obtained independent components may be irrelevant [35].

3.2 Experimental Results

In this section, we aim to obtain a new representation of the data, having a lower dimensionality but with more informative features. Several experiments were performed to compare multiple DR techniques in terms of accuracy, F1-score, and standard deviation. The classification task in these experiments have been carried out using DT and kNN classifiers. We compare 22 DR methods, namely PCA, FA, CFA, MDS, LDA, ISO, SM, PPCA, LE, LLE, LIM, HLLE, LTSA, KPCA, GDA, NPE, LPP, DM, LLC, MC, LMNN, and ICA. Fig. 2.2 shows the DR techniques that are utilized in the designed IDS.

3.2.1 Experimental Setting

The SCADA system records the network flow data in the water storage system, which are captured via a serial port data logger. The recorded data has 200,000 samples.

The recorded network traffic data consists of 24 unique features, as shown in Table 2.2, that are used to detect malicious activities. The network traffic data is recoded from MODBUS traffic with RS-232 connection, in which it is one byte long and each server has a unique device address.

Read and write commands/responses, which have a fixed length for each system, are supported to observe an intrusion to the system. When there is no intrusion, the error rate should be low and constant but when the system experience a denial-of-service attack the rates are predictable to rise. When no error is detected (during the normal state), the response function code matches the command function code and the existence of an error the response sub-function code is changed to command function code plus a value of 0X80.

3.2.2 Results Analysis

We evaluated the performance of 22 dimensionality reduction methods and divided the train and test data on the basis of k-fold cross-validation approach, using k = 10. Data is divided into partitions as train/test based on "K". Here, K refers to any integer while fold is to a partition (or iteration). Each model is trained on K-1 partitions and tested on K-th partition of data, and then, the results obtained through testing over ten folds are averaged and reported.

In general, Fig. 3.1 and 3.2 show the impact of dimensionality reduction on the performance of intrusion detection system. Fig. 3.1 shows the accuracy obtained by each classifier without performing dimensioanlty reduction, while Fig. 3.2 shows the accuracy obtained by each classifier after applying dimensioanlity reduction (PPCA). Each bar shows the accuracy obtained through each cross-validation fold. In general, kNN classifier outperforms the decision tree classifier in terms of accuracy. Considering the original dataset, kNN recorded an average accuracy of 87.89% and DT results in an average accuracy of 87.92%. Besides, when DR is applied, kNN and DT achieve an averaged accuracy 99.96% and 99.88%, respectively.

In Fig. 3.3, it is apparent that performance of the PPCA method is consistently and significantly higher when combined with kNN or DT compared to other DR methods, and it reduces the dimensionality of the feature space to 10 features. PPCA method obtains the highest performance in terms of accuracy and F1-score

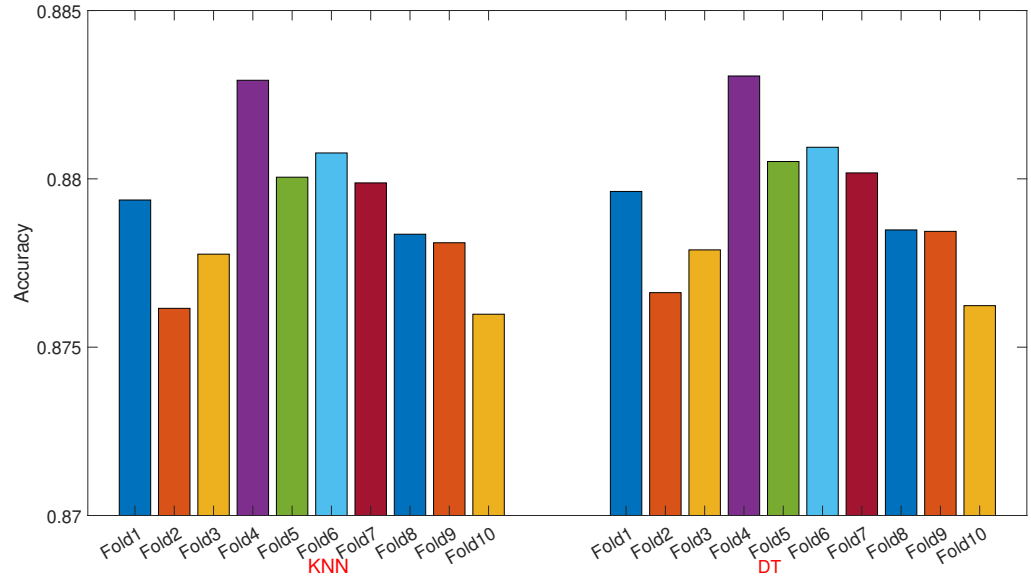

Figure 3.1 – Accuracy obtained by each classifier at each cross-validation fold without performing dimensionality reduction.

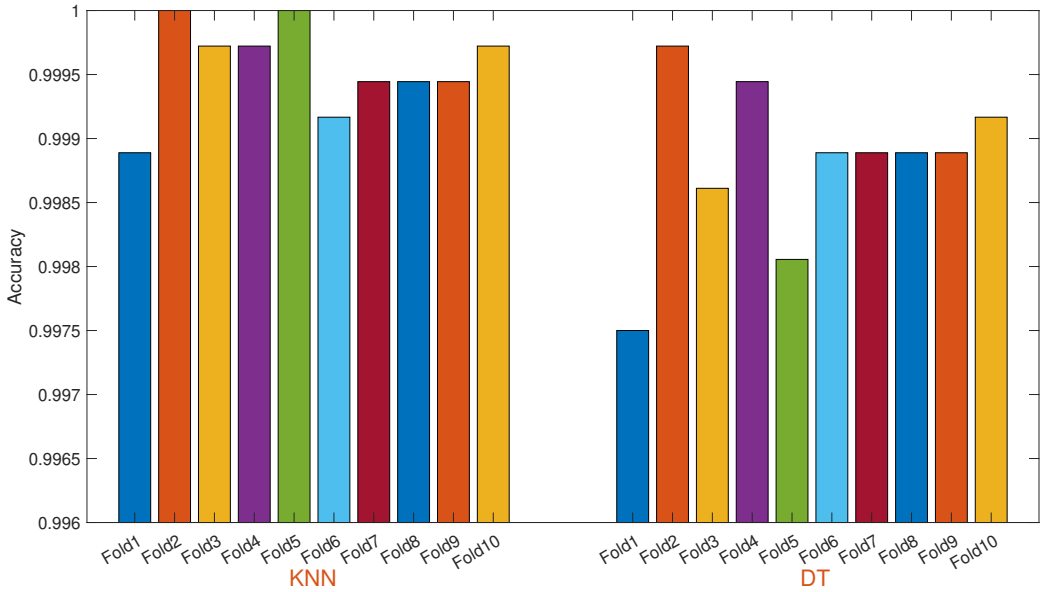

Figure 3.2 – Accuracy obtained by each classifier at each cross-validation fold after performing dimensionality reduction (PPCA).

compared to other DR methods. Many dimensionality reduction methods perform reasonably well, and their performance is relatively stable across a range of included low-dimensional components. In terms of the accuracy measure and the kNN classifier, Fig. 3.3(a) shows that PPCA method has outperformed the other methods, as shown lightest yellow color. This is while MC and LLC methods are ranked second and third, albeit with a slight difference. Furthermore, LDA, CFA, KPCA, ICA, LTSA, HLLE, ISO, LE, SM, LMNN, GDA, LLE, MDS, FA, LIM, DM, NPE, LPP and PCA methods are ranked from fourth to 22-th, respectively. MC and LLC methods have a desirable performance with an average accuracy from 0.982 to 0.986. LPP and PCA have failed to improve the classification performance using the kNN classifier that results in an accuracy lower than 70%. Considering the results of DT accuracy in Fig. 3.3(b), PPCA is ranked as first, and it is followed by LDA and MC that are ranked as second and third, with a slight difference. ISO, ICA, CFA, MDS, KPCA, LTSA, HLLE, LE, SM, GDA, FA, PCA, LMNN, LIM, LLE, and LLC methods are ranked from forth to 19-th. NPE and DM methods share the 20-th rank as they show equal performances. Lastly, LPP was ranked as the last technique, as it recorded less than 67%.

In addition to accuracy, Fig. 3.3(c) and (d) represent the F1-score performance for the kNN and DT classifiers, respectively they also indicate that PPCA yields the highest F1-score, almost 99%, and is ranked as the best. Furthermore, MC, LLC, LDA, and CFA methods result in an average F1-score between 98% and 97%, when combined with KNN, and can be considered as the second-best algorithms. Moreover, ISO, SM, LTSA, HLLE, LLE, LE, LMNN, LIM, KPCA, MDS, and FA methods are ranked from six to 16-th with an average of 84% to 94%. On the other hand, Fig. 3.3 (d) shows that when the DT classifier is employed, those methods that come after PPCA are: LDA, MC, PCA, ISO, SM, CFA, FA, LTSA, HLLE, LE, LLE, LIM, LMNN, and LLC that are ranked from second to 15-th place. Both NPE and DM methods share the 17-th place, when the KNN classier is used with average of 82% .

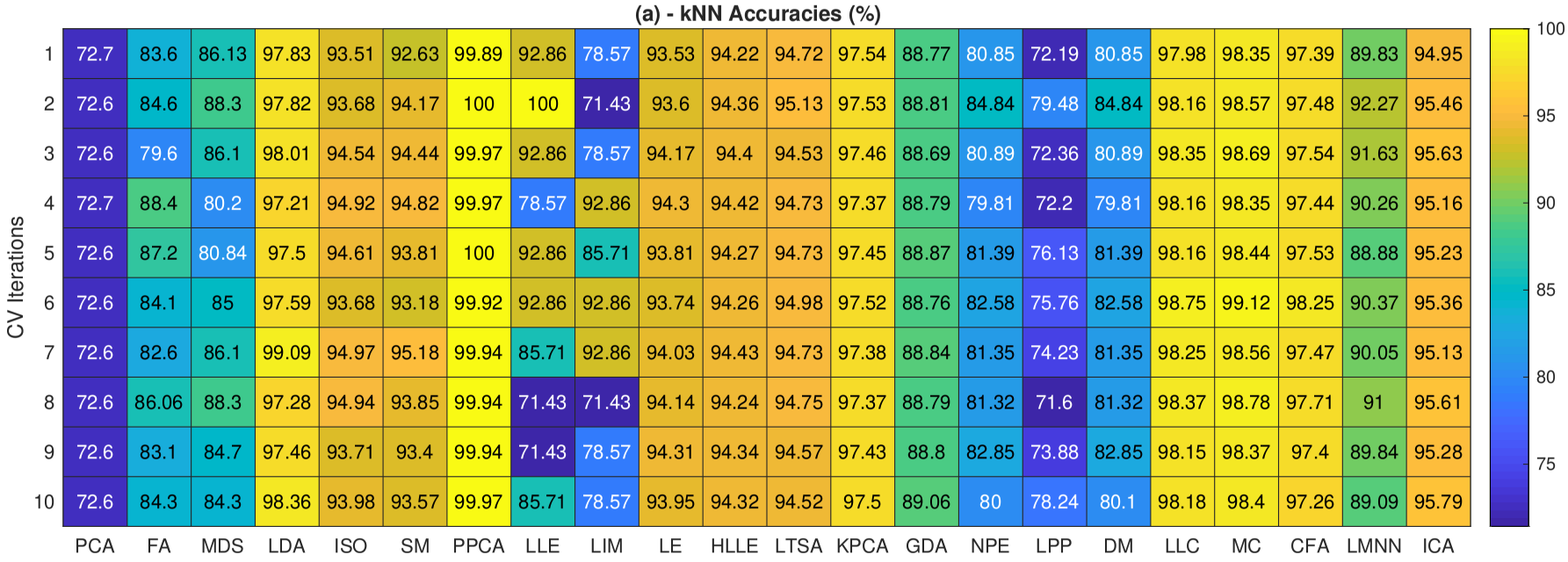

Figure 3.3 – (a) Obtained accuracy through each combination of the DR techniques with kNN.

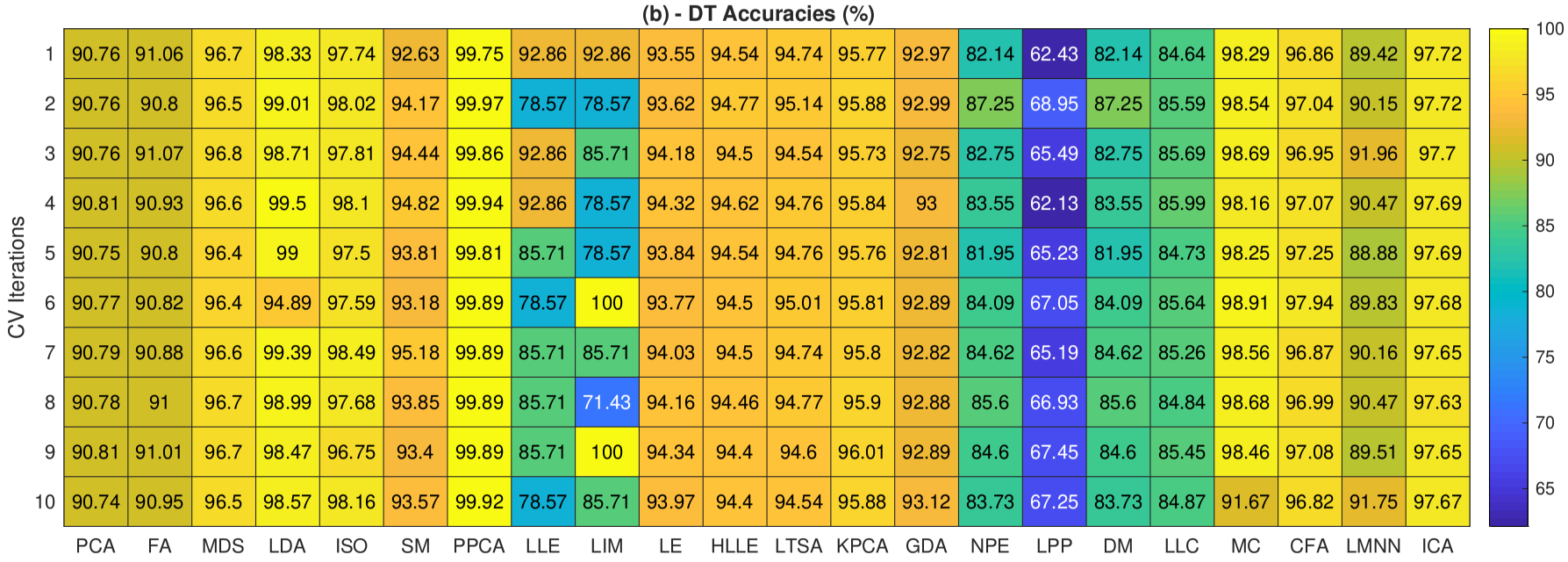

Figure 3.3 – (b) Obtained accuracy through each combination of the DR techniques with DT.

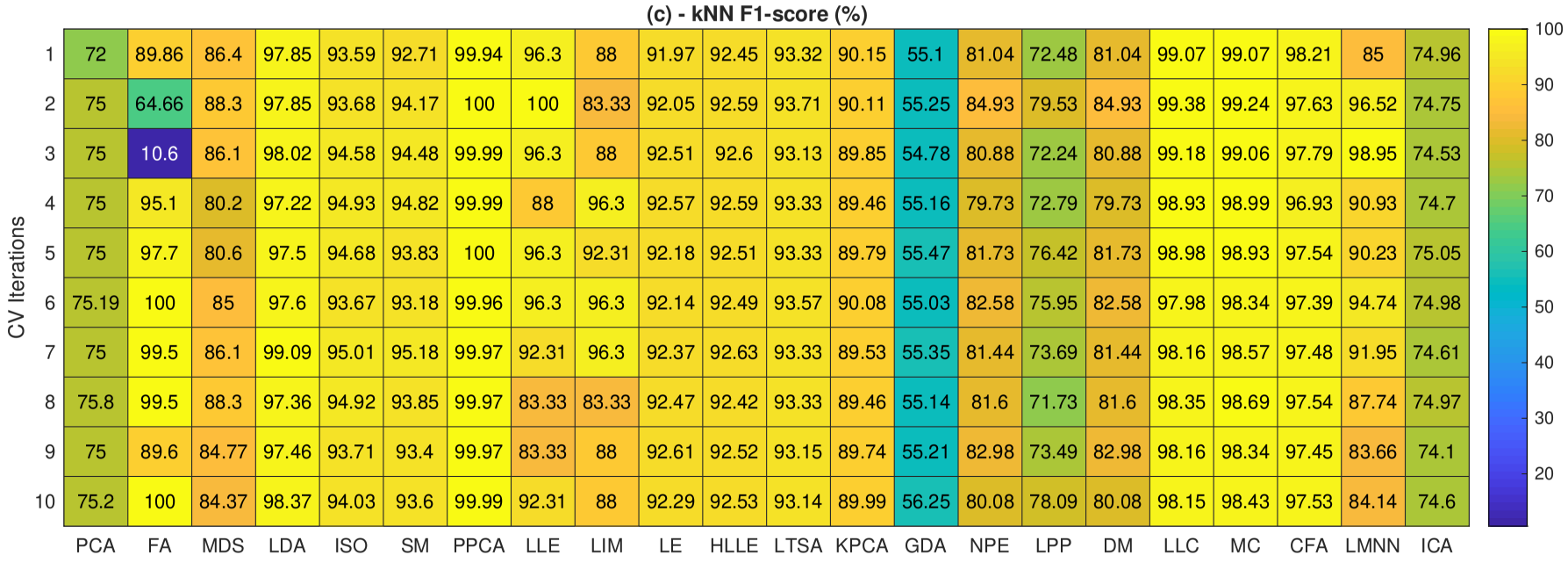

Figure 3.3 – (c) Obtained F1-scores through each combination of the DR techniques with kNN.

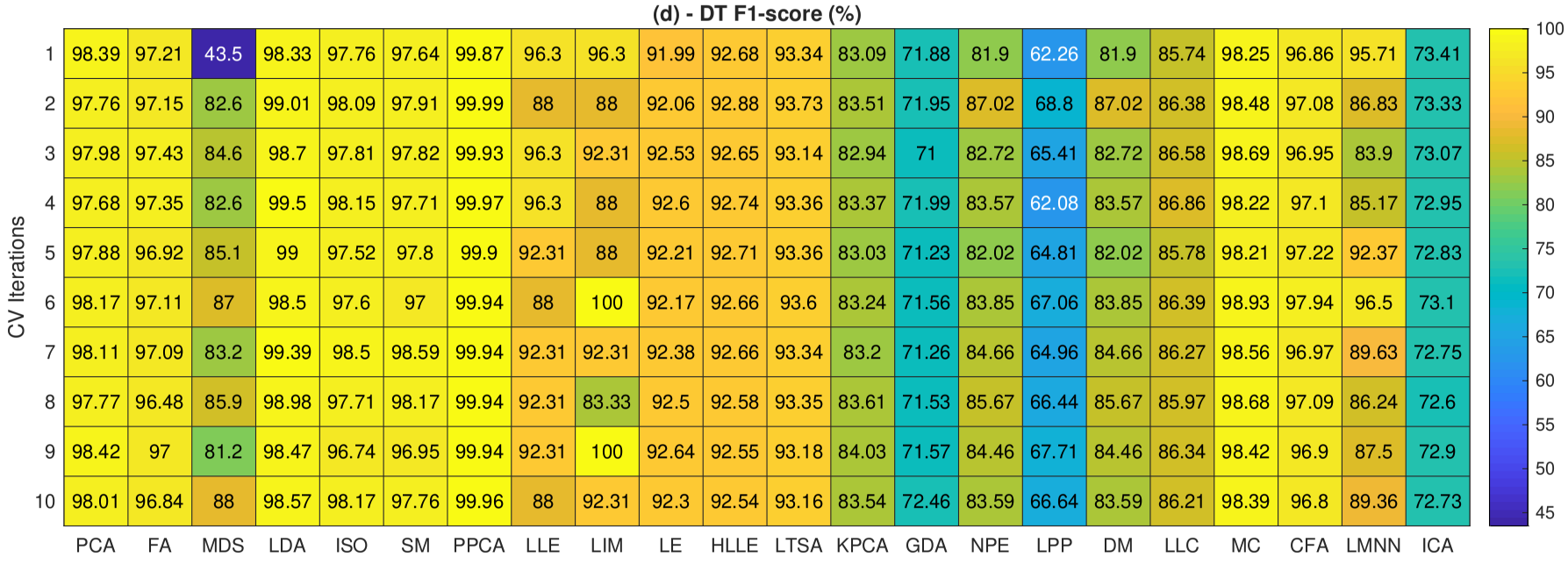

Figure 3.3 – (d) Obtained F1-scores through each combination of the DR techniques with DT.

While using DT classifier NPE and DM are ranked 16th, then KPCA and MDS are ranked 17 to 18th, respectively. Lastly, using the KNN classifer, PCA, ICA, LPP, and GDA methods are ranked as the last methods, whereas using the DT classifier, ICA, GDA, and LPP are ranked from 20-th to the last. Both KPCA and GDA methods produce a small vector of dimension two with an average accuracy score between 0.90 to 0.96.

Generally, GDA, LPP, and ICA methods are not sensitive to the choice of classifiers as they result in lower F1-scores than others in average of 50% and 70%. PPCA, MC, LLC, LDA, and CFA are more compatible with KNN, while the rest of the dimensionality reduction methods like PCA, ISO, SM, CFA and FA are suggested to be used with the DT classifier due to higher performance tange 97%. In general, PPCA outperforms all the competitors and results in the maximum accuracy when coupled with KNN or DT classifier. Similarly, GDA worsens the F1-score; however, the stability is improved when it is used with the DT classifier.

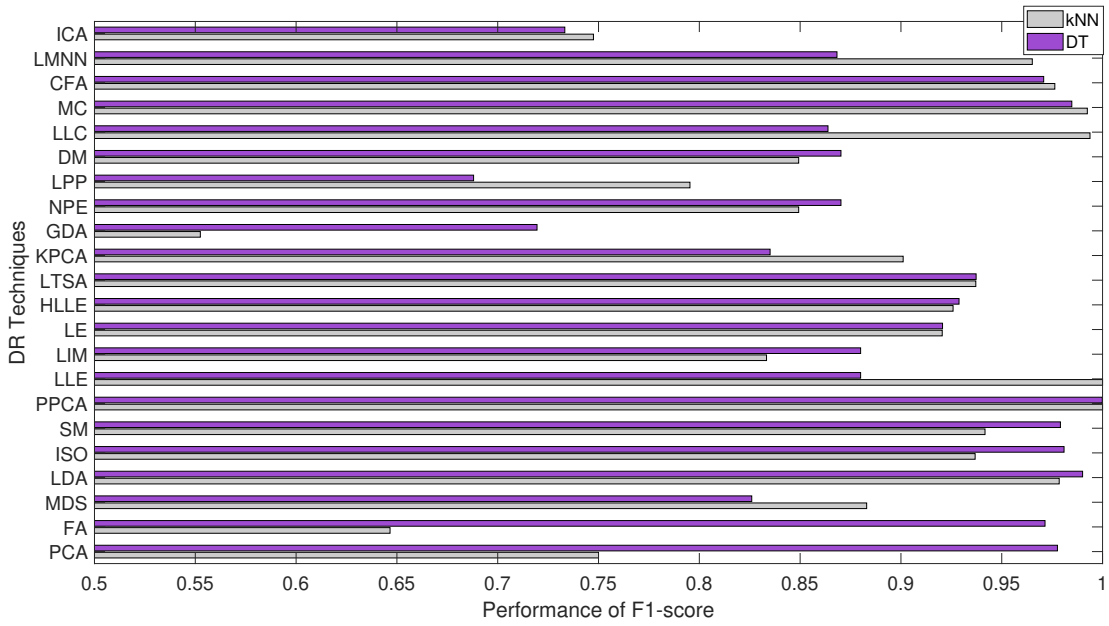

Figure 3.4 – F1-score measures attained through each DR technique along with the DT and KNN classifiers.

Fig. 3.4 indicates the relatively higher F1-score performance achieved by the DR methods in comparison with individual DR algorithms. The results demonstrate better performance (more closer to 1) when the DT classifier is used instead of the KNN classifier for most DR methods, for instance, 12 methods (PCA, FA, LDA, ISO, SM, LIM, LE, HLLE, LTSA, GDA, NPE, and DM) scored higher in the DT classifier while 9 methods (MDS, LLE, KPCA, LPP, LLC, MC, CFA, LMNN, and ICA) in KNN classifier. The best performance in terms of the F1-score is almost 100% that is obtained by both the KNN and DT classifiers,in combination with PPCA and using the dimensionality size of 10 feature.

3.3 Summary

A scheme has been designed for classifying cyber-attacks to study the effect of dimensionality reduction on the classification performance. A SCADA system of a water storage tank used in this study. This cyber-physical system undergoes multiple cyber-attacks in our study, for which we design the cyber-attack identification scheme. The designed scheme leverages 22 advanced dimensionality reduction techniques that are couple with two classifiers. This hybrid scheme enables a comparative study on the impact of DR methods and their compatibility with the selected classifiers. These algorithms are compared in terms of accuracy, F1-score, and standard deviation. The conducted analysis ranks all of these DR methods and finds the best combination for the optimal classification accuracy in this cyber-physical water system.

Chapter 4

Effect of Feature Selection on Intrusion Detection

The quality of data in a data-driven process is usually affected by various factors [28, 29]. Feature selection, also known as variable selection or attribute selection, is a common technique for improving the data quality. This process obtains a subset of relevant features and eliminates the irrelevant and redundant features from the original data. Such improvement in data quality will in turn enhance the performance of data-driven modules such as change detectors [13, 54] and classifiers [15, 16] in the system. FS methods have different criteria, such as their variance, entropy, and ability to preserve local similarity, which results in different correlation and consistency.

4.1 Review of Feature Selection Methods

Commonly, feature selection and dimensionality reduction are banded together, as both methods are used for reducing the number of features in a dataset. But the important difference between both methods is that feature selection is simply selecting and excluding given features without changing them, and dimensionality reduction transforms features into a lower dimension. Feature selection can improve the accuracy of the model, by removing features with missing values, removing features low variance, removing highly correlated features, and feature selection using SelectFromModel. Most feature selection methods are divided into three major categories: (i) filter-based: Filter methods are generally used as a preprocessing step, as they choose intrinsic properties of the features measured via univariate statistics instead

of cross-validation performance. When dealing with high-dimensional data, it is computationally cheaper to use filter methods. (ii) wrapper-based: this method is based on a specific machine learning algorithm to find optimal features, it finds approach by evaluating all the possible combinations of features against the evaluation criterion. Feature subset is obtained by checking its usefulness in classification, as the estimated predictive accuracy is typically considered to be the most important indicator of relevance for attributes. The wrapper methods usually perform in higher predictive accuracy than filter methods. (iii) embedded: is an iterative method, in which the selection process is utilized into the learning of the classifier that takes care of each iteration of the model training process and carefully extracts those features which contribute the most to the training for a particular iteration. Most of these methods can perform two processes: ranking and subset selection (sometimes they are performed sequentially): the importance of each individual feature is estimated, commonly by disregarding potential interactions among the elements of the joint set, then, the final subset of features to be selected is provided.

While feature selection techniques often operate singularly and are not combined with other feature selection algorithms, it is also possible to use these techniques in combination. By doing so, one can use a simple approach such as the majority of votes to aggregate the results of these techniques. However, this approach will be most advantageous when the selected algorithms employ completely different methods (e.g., manifold learning, cluster analysis, and mutual information) to capture the distribution of the feature space. This will extend the flexibility of the feature selection process against various distribution types and data structures.

4.1.1 Infinite Feature Selection

The infinite feature selection (InfFS) is a filter-based technique that models the feature space using graphs. In this process, each graph node corresponds to a feature, and edges connecting these nodes represent pair-wise relationships between features.

Weighted edges of this graphical model codify the independence between two feature distributions. A path on this graph then shows a subset of features. The convergence properties of the power series of matrices and Markov chain fundamentals are then used to evaluate the paths that contain certain features. InfFS determines a final score that shows the best feature candidate by ranking in a descent order [55].

4.1.2 Infinite Latent Feature Selection

Similar to InfFS, Infinite latent feature selection (ILFS) is a probabilistic technique that models the feature space using a graph-based approach that considers all the possible subsets of features during the ranking process. However, ILFS models the relevancy between features as a latent variable in a generative process, which is inspired by the probabilistic latent semantic analysis. This enables the algorithm to investigate the feature importance upon the injection of a feature into an arbitrary set of cues [56].

4.1.3 Evolutionary Computation Feature Selection

Evolutionary computation (ECFS) has the ability to search simultaneously within a set of possible solutions to find the optimal and effective solution set, by iteratively trying to improve the feature subset with regard to a given measure of quality. An outline of three steps of the EC algorithm are as follow: 1) initialization, where the population of solutions is initialized randomly; 2) evaluation of each solution in the population for fitness value; 3) iteratively generating a new population until the termination criteria (e.g., could be the maximum number of iterations or finding the optimal set of features that maximizes classification accuracy) are met [57].

4.1.4 Relief Feature Selection

Relief Feature Selection (ReliefFS) calculates a proxy statistic (referred to as feature weights) for each feature that can be used to estimate feature quality or relevance to the target concept. Relief is supplanted by ReliefFS which relies on a number of neighbors user parameter k that specifies the use of k nearest hits and k nearest misses in the scoring update for each target instance. ReliefFS finds k nearest misses from each class and averages the weight update based on the prior probability of each class [58].

4.1.5 Mutual Information

Mutual information (MutlnfFS) is a measure of dependency between two (possibly multi-dimensional) random variables that shows how much knowing the value of one variable reduces the uncertainty on the others. MI is also able to capture non–linear dependencies and is invariant under invertible and differentiable transformations of the random variables, in which it has been used as a score in the filter methods. The selected features will be those with top mutual information w.r.t. classes [59].

4.1.6 Maximum Relevance and Minimum Redundancy

In the Maximum Relevance and Minimum Redundancy (mRMR) method, each feature can be ranked based on its relevance to the target variable, and the ranking process is able to consider the redundancy within the selected features. The best feature is defined as one that can effectively reduce the redundant features while keeping the relevant features for the model [60].

4.1.7 Feature Selection via Concave Minimization

Feature Selection via Concave Minimization (FSV) is considered as a wrapper method, in which subsets of features are sampled, evaluated, and finally kept as the final out-

put. FSV generates a separating plane by minimizing a weighted sum of the distances of misclassified points to two parallel planes that bound the sets, and determines the separating plane midway between the set of misclassified points [61].

4.1.8 Laplacian Score

Laplacian score (Laplacian) is based on two data points that are probably related to the same topic if they are close to each other, in which it is based on the Laplacian Eigenmaps and Locality Preserving Projection. For each feature, the Laplacian score is computed to reflect its locality geometric structure so features that are consistent with the Gaussian Laplacian and with small weighted variance are selected [62].

4.1.9 Multi-Cluster Feature Selection

Multi-Cluster Feature Selection (MCFS) uses a multi-cluster structure that is defined to measure the correlations between different features without the use of label information (unsupervised feature selection). Recently, the spectral clustering structure of the data shows a significant interest, in which data points are structured using the top eigenvectors of graph Laplacian (manifold learning) and find the subset selection using L1-regularized models [63].

4.1.10 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is basically a recursive process that ranks features according to some measure of their importance. The less relevant feature is removed iteratively since it has the least effect on the classification. Therefore, RFE aims to eliminate dependencies and collinearity that may exist in the model. For high correlated features and large datasets the relative importance of each feature can change substantially when analyzed over a different subset of features during the stepwise elimination process, in which recursion is used [64].

4.1.11 L0-Norm

Norms are a way to measure size or length in higher dimensions. L0-norm is the most direct and ideal scheme of feature selection that is difficult to optimize so L0-norm can balance the training error against the number of non-zero features [65]. L0-Norm penalizes features by which the regularization and parallel parameter estimation processes become more complicated. L0-norm solves the L0 penalty problem by selecting non-zero coefficients and regularization parameters simultaneously. In addition, it finds an estimated solution for this penalty problem.

4.1.12 Fisher Score

Fisher score finds a subset of feature, which selects the top-ranked features with large scores. The score of each feature is computed independently by the heuristic algorithm. The algorithm fails to select features that have low individual scores but a very high score when they are combined together as a whole [66].

4.2 Experimental Results

This section analyses the obtained results in terms of accuracy and standard deviation and compares it with twelve different FS techniques. Fig. 2.1 shows all the feature selection techniques that were used in our approach.

4.2.1 Experimental Setting

The water storage tank system generates network flow records that are captured with a serial port data logger which include 200,000 samples. 19503 of these samples correspond to the normal state (i.e., class 0), and the rest of the samples are collected when the system was under attack. Classes 1 to 7 in Table 2.1 have 1198, 1457, 209, 410, 155, 135, and 4132 samples, respectively. In order to detect malicious activities

in the Water Storage System, features are divided into network traffic features and payload content features. The former gives information regarding the communications within the SCADA network system, while the latter describes the current state of different components of the SCADA system. The developed dataset consists of 24 unique features (i.e., 8 payload and 16 network traffic features), as shown in Table 2.2.

The dataset described in this chapter used MODBUS traffic from RS-232 connection, in which it's one byte long with each server having a unique device address. The water storage tank holds approximately two liters of water that consist of a relief valve to drain water from the tank, a pump to add water to the tank, and a meter to measure the percentage of water level. In addition to the on/off control scheme to maintain the water level between high (H) and low (L) set-points, an alarm is turned on when the water level is above high alarm set-point (HH) or below the low alarm set-point (LL).

In order to log the data and inject attacks, a bump-in-the-wire method is used. The device implementation is conducted using C programming and VMware virtual machine. Two RS-232 serial ports are included in the virtual machine that are connected to a USB-to-serial converter. The programmed software monitors serial ports for traffic. Any detected traffic is then timestamped and saved in a log file. Furthermore, the software incorporated hooks to inject, delay, drop, and alter network traffic to facilitate the attacks.

4.2.2 Results Analysis

Fig. 4.1 shows the performance measure in terms of accuracy through 10-fold cross-validation and by resorting to feature selection techniques along with the kNN and DT classifiers. Without the feature selection is being operated on the dataset, the DT classifier displays better results than the kNN classifier (i.e, slightly improve). As observed in the figure that feature selection improves dataset. However, after

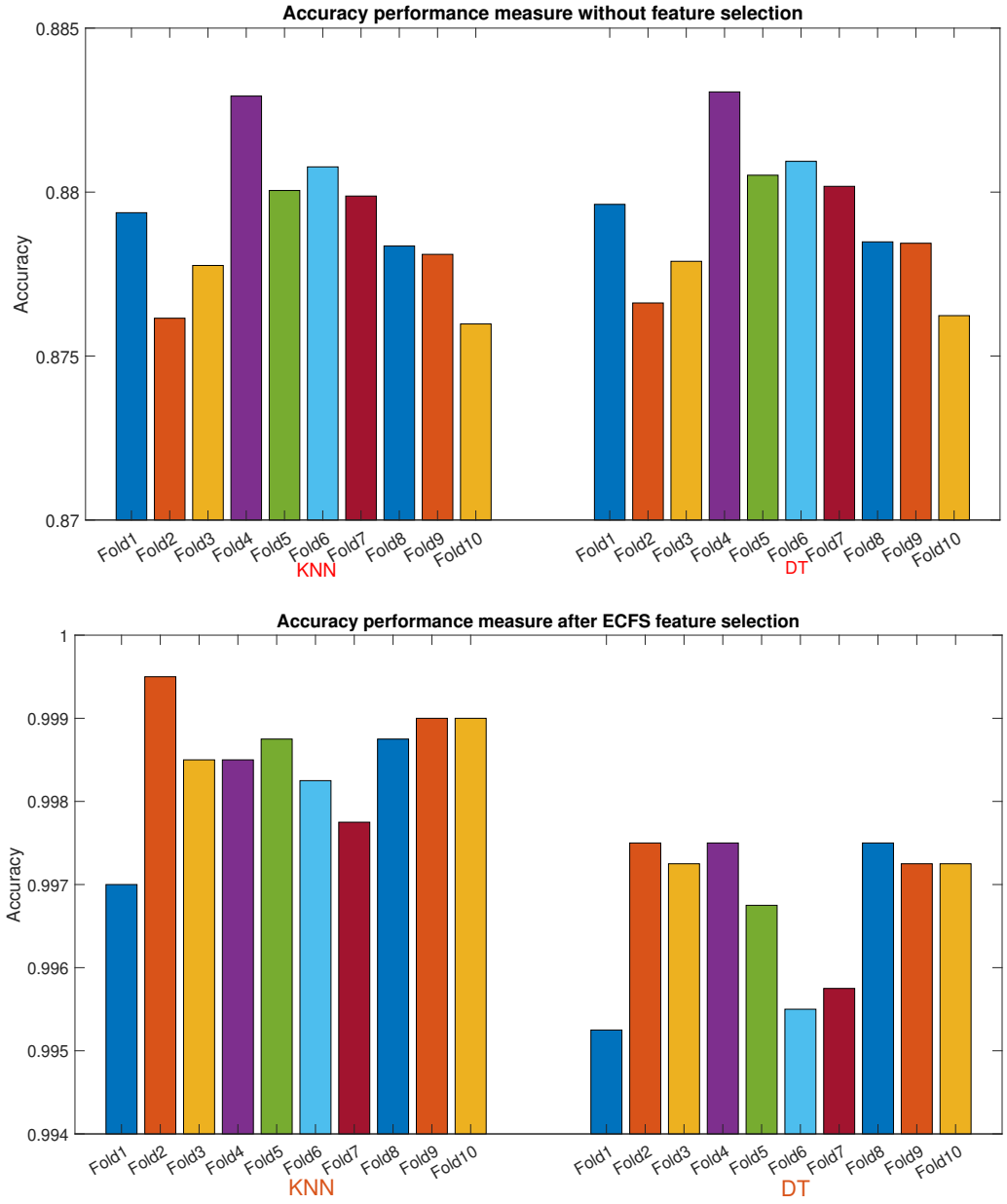

Figure 4.1 – Classification accuracy obtained by DT and kNN without performing feature selection (top panel) and after performing feature selection (bottom panel). Each bar shows the accuracy obtained at each fold of the cross-validation.

the ECFS feature selection, the kNN classifier shows a satisfying improvement and recorded a higher average accuracy than the DT classifier.

In Fig. 4.2, 4.3, 4.4 and 4.5, it is conspicuous that ECFS, InfFS, and ILFS methods show higher performance in terms of accuracy and F1-score. The accuracy performance attained by the kNN classifier is demonstrated in Fig 4.2, in which the ECFS method illustrates the best results among all FS methods, and its average accuracy is approximately 99.85%. Furthermore, ILFS and InfFS are ranked second and third with the average accuracy of 99.7%, and it is likewise considering the F1-score performance on kNN, as shown in Figure 4.4. Moreover, mRMR, ReliefF, RFE, Fisher, L0-Norm, MCFS, MutInFS, FSV, and Laplacian methods are ranked from forth to 12-th, while their accuracy performance falls between 82% and 98% on the kNN classifier.

The presentation of accuracy performance using the DT classifier is shown in Fig. 4.3 in which InfFS method has outperformed the other FS methods with an accuracy measure of 99.7%. ECFS and ILFS methods come in the second and third ranks with an average of 99.6%. The rest of the methods are ranked from forth to twelve12-th and sorted as: Fisher, RFE, L0-norm, mRMR, ReliefFS, MutInfFS, FSV, MCFS, and Laplacian. ECFS, ILFS, and InfFS methods result in the best accuracy compared to the other nine FS methods. In addition, the Laplacian method is less likely to be sensitive to the choice of classifiers.

To study the F1-score on the kNN and DT classifiers, Fig. 4.4 and 4.5 illustrates the results of the twelve utilized FS techniques. Considering the results of the kNN classifier, ECFS recorded the highest F1-score when combined with the kNN classifiers in Fig. 4.4 and Fig. 4.5, respectively. ILFS and InfFS maintain their second and third ranks when coupled with kNN classifier. Using DT classifier, however, changes InfFS recorded the highest F1-score with a score of 0.997 (99.7%), ECFS and ILFS ranks to second and third, respectively. Moreover, in respect of the kNN classifier, mRMR, RFE, Fisher, L0-Norm, MCFS, FSV, ReliefF, Laplacian, and MutInfFS methods are

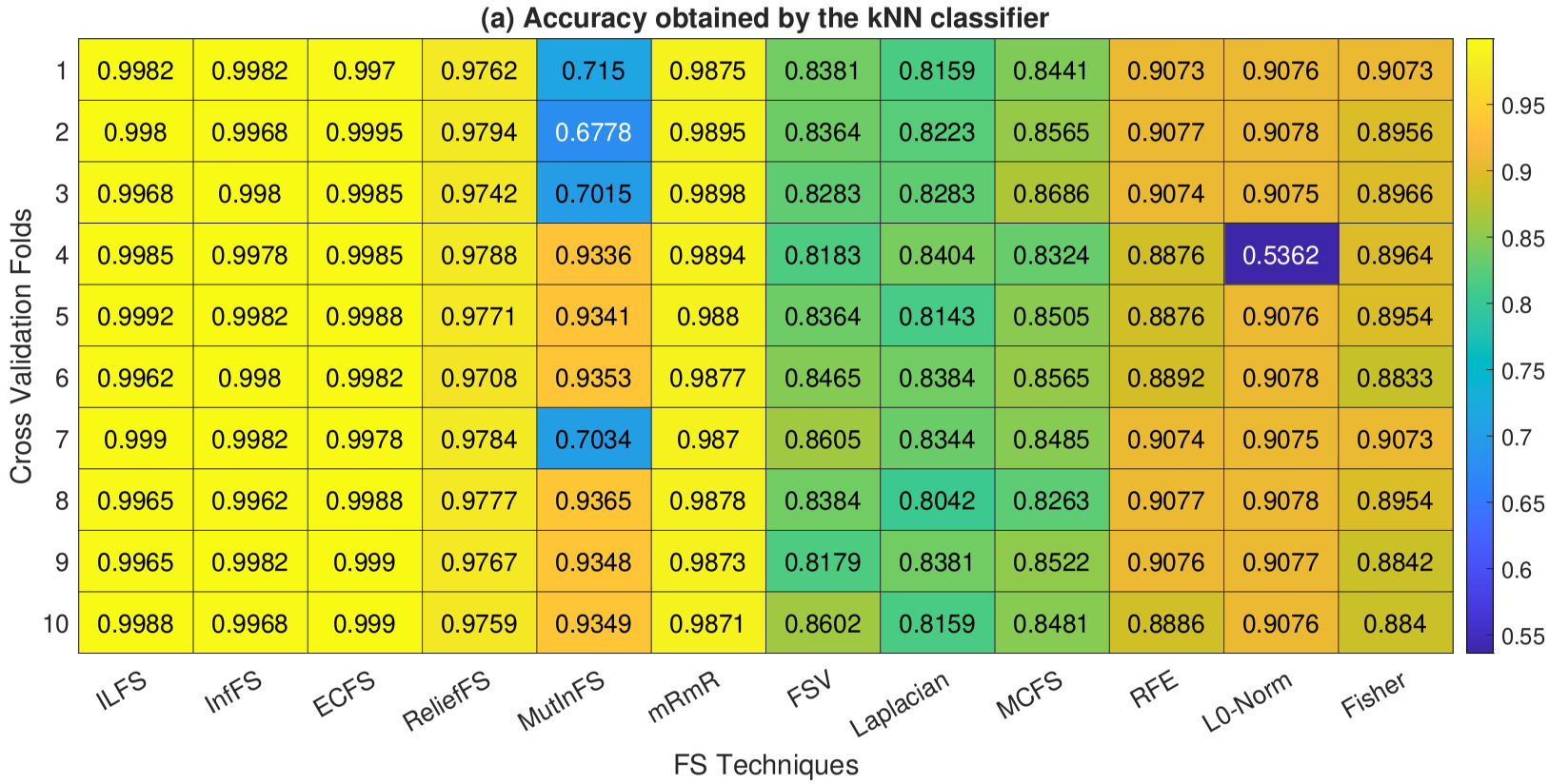

Figure 4.2 – Accuracy measures obtained through each FS method in different iterations of cross-validation by resorting to the kNN classifier.

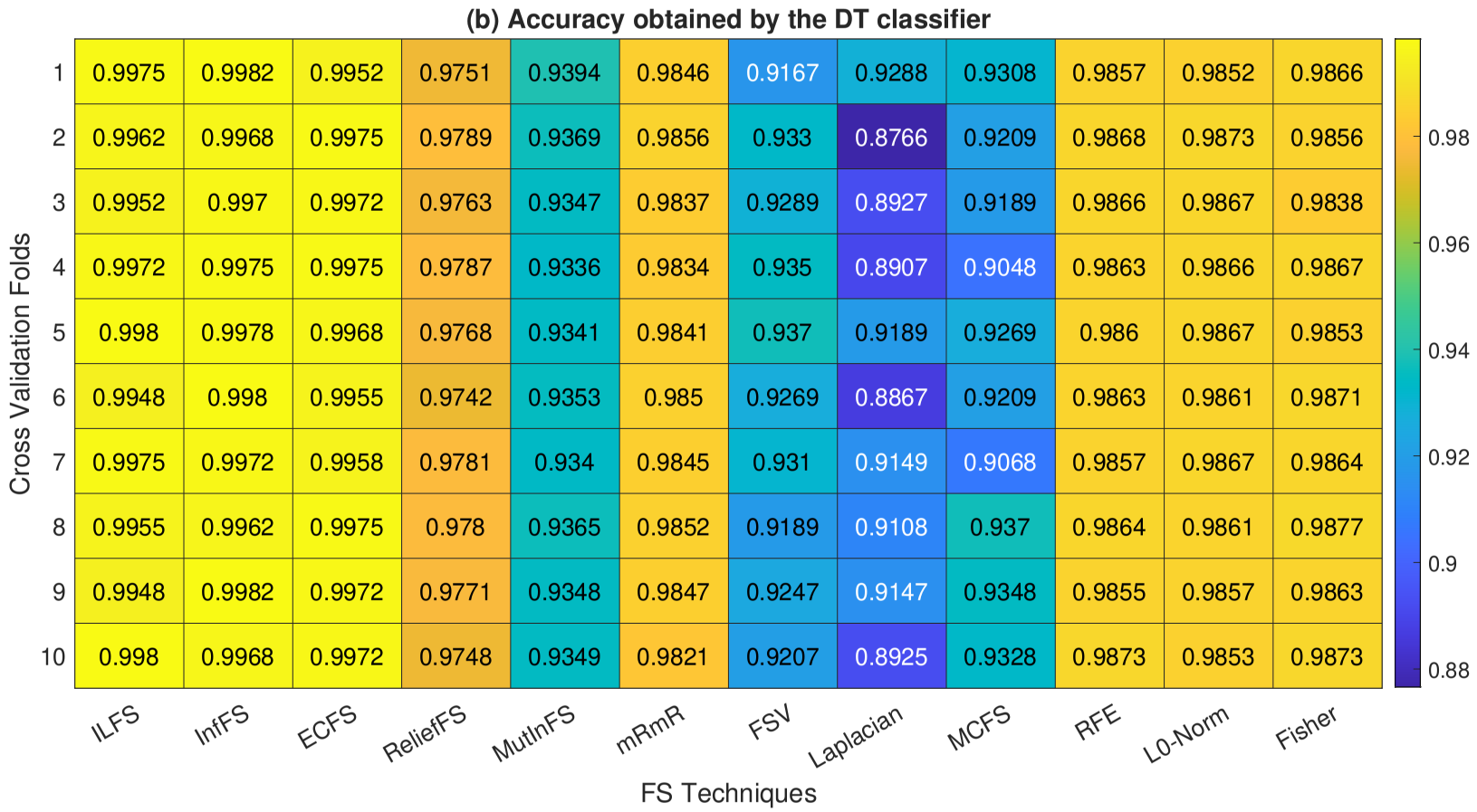

Figure 4.3 – Accuracy measures obtained through each FS method in different iterations of cross-validation by resorting to the DT classifier.

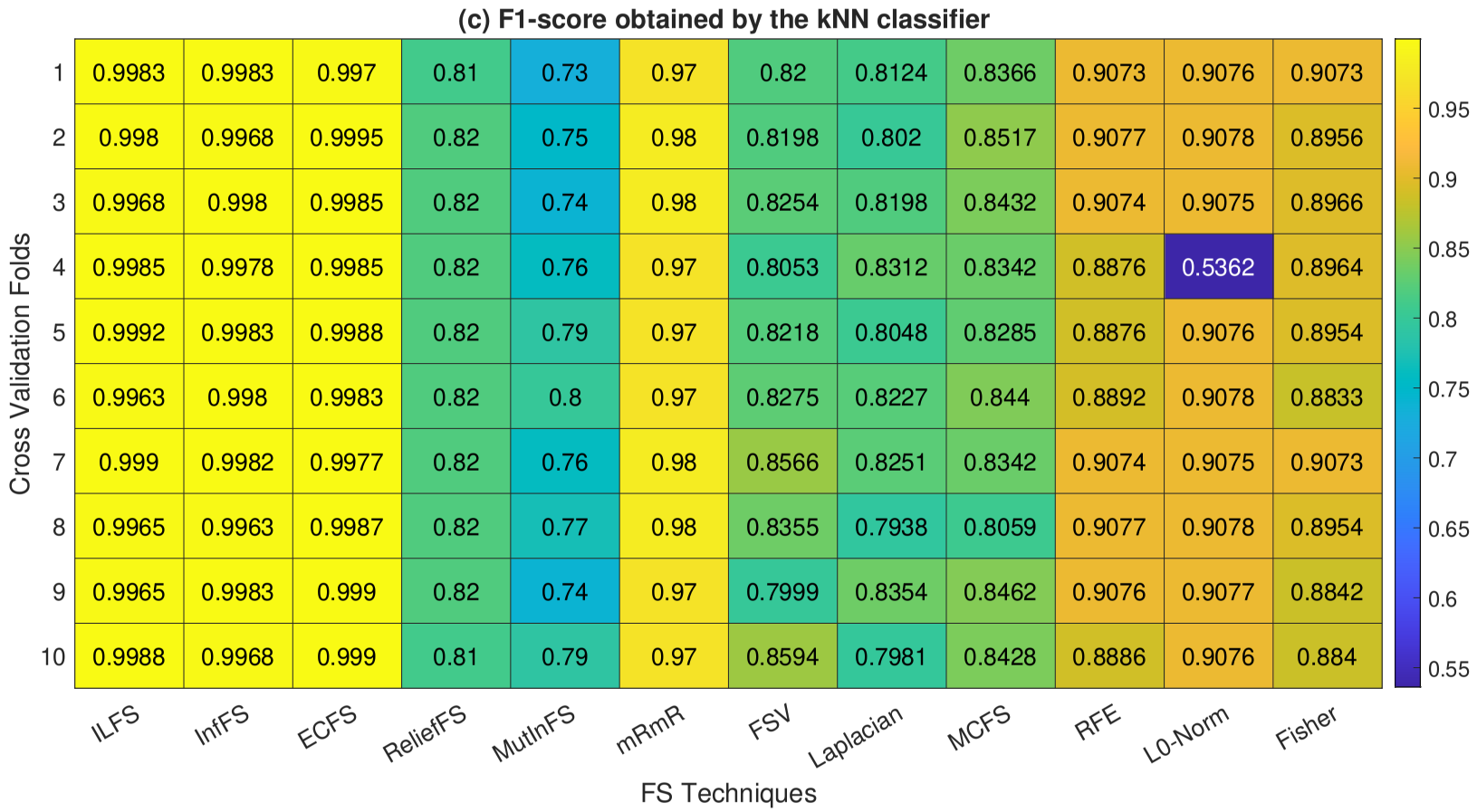

Figure 4.4 – F1-score measures obtained through each FS method in different iterations of cross-validation by resorting to the kNN classifier.

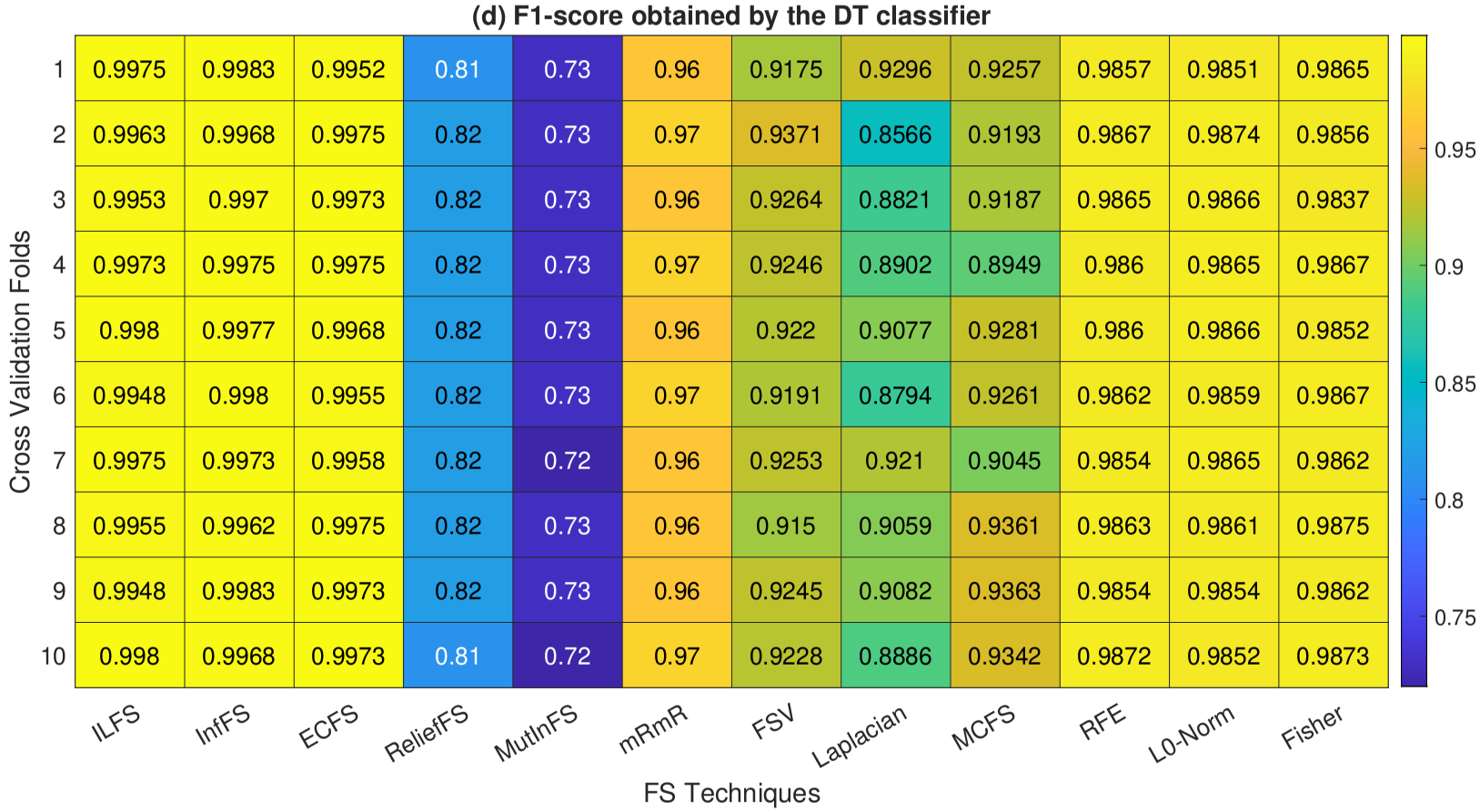

Figure 4.5 – F1-score measures obtained through each FS method in different iterations of cross-validation by resorting to the DT classifier.

ranked from fourth to 12-th, respectively. Considering the DT classifier, FS methods are ranked in the following order from fourth to the 12-th rank: Fisher, RFE, L0-norm, mRMR, FSV, MCFS, laplacian, REliefF, MutInfFS. ECFS, ILFS, and InfFS methods are more stable and always improve the F1-score. MutInFS method has failed to improve the F1-score performance, and Laplacian failed to improve the accuracy performance.

Fig.4.6 shows the F1-score performance obtained through ECFS, ILFS, and InfFS methods, which are more compatible with the kNN classifier. This is while the rest of FS methods are suggested to be used along with the DT classifier. In general, ECFS performs better than other FS techniques and results in the maximum accuracy when coupled with kNN that is about 99.85%. In addition, MutInFS worsen the F1-score; however, stability is improved when it is used with kNN which is about 76.3%. RFE, L0-norm, and Fisher techniques result in a stable and a slight difference in F1-score when coupled with the DT classifier, in which it scores close to 98.6%. These results have shown that feature selection method are effective and robust in the classification of SCADA datasets.

4.3 Feature Analysis

Considering the FS outputs for all the studied FS methods, Fig.4.7 illustrates the importance of each feature w.r.t. the number of times it is selected by FS methods. Based on the results shown in this figure, it can be inferred that the most important features are the response address and time (number 2 and 23 in Table 2.2). The second most important features are features 9 and 15, which are response read function and HH, respectively. The third group of important features are command address, command memory, and command length (numbers 1, 3, and 12 in Table 2.2). These seven features are selected more than six percent of the times and they are believed to be most effective on the detection accuracy. This while the least informative features

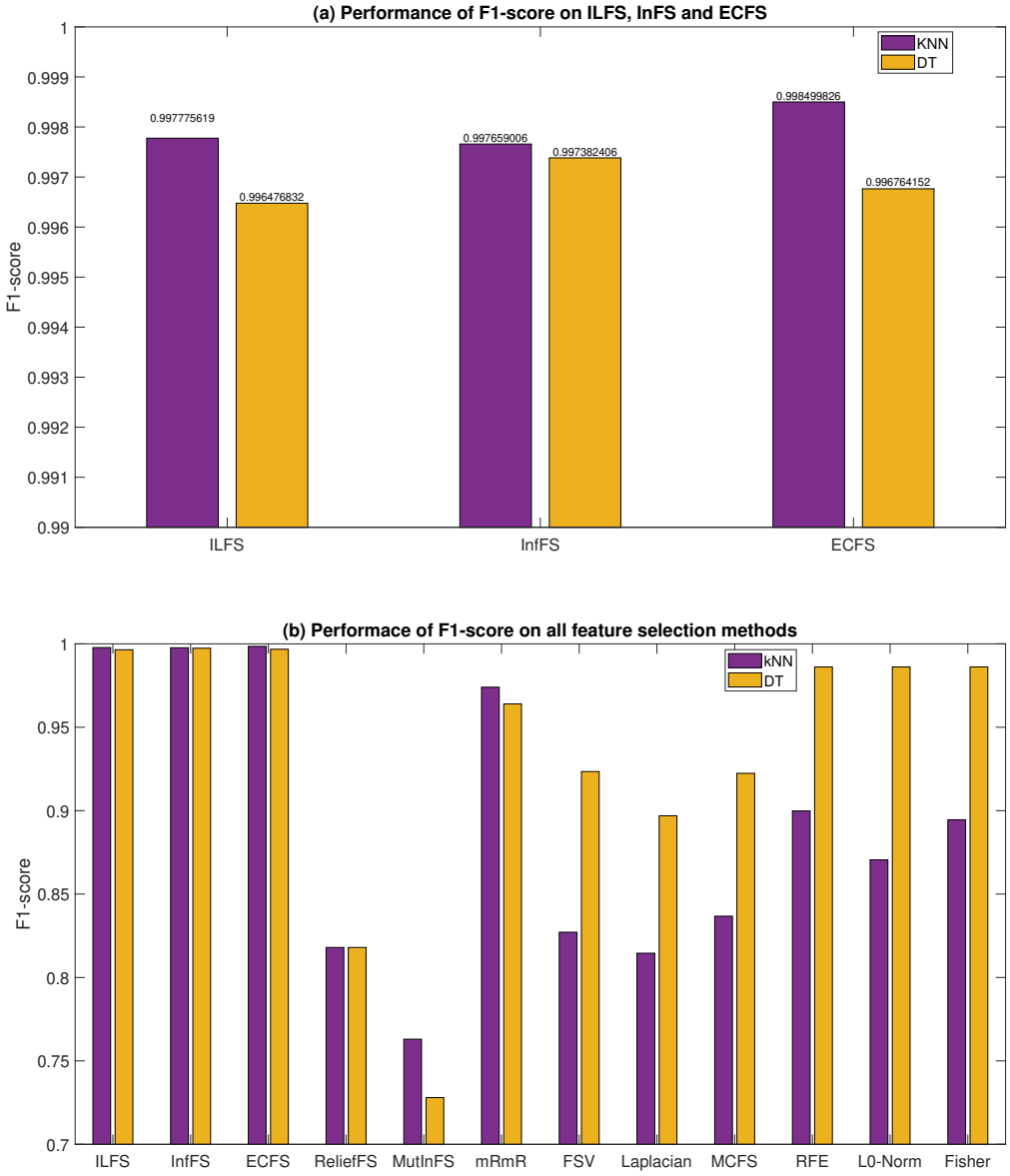

Figure 4.6 – Averaged F1-score attained through each FS method along with DT and KNN over ten cross-validation folds.

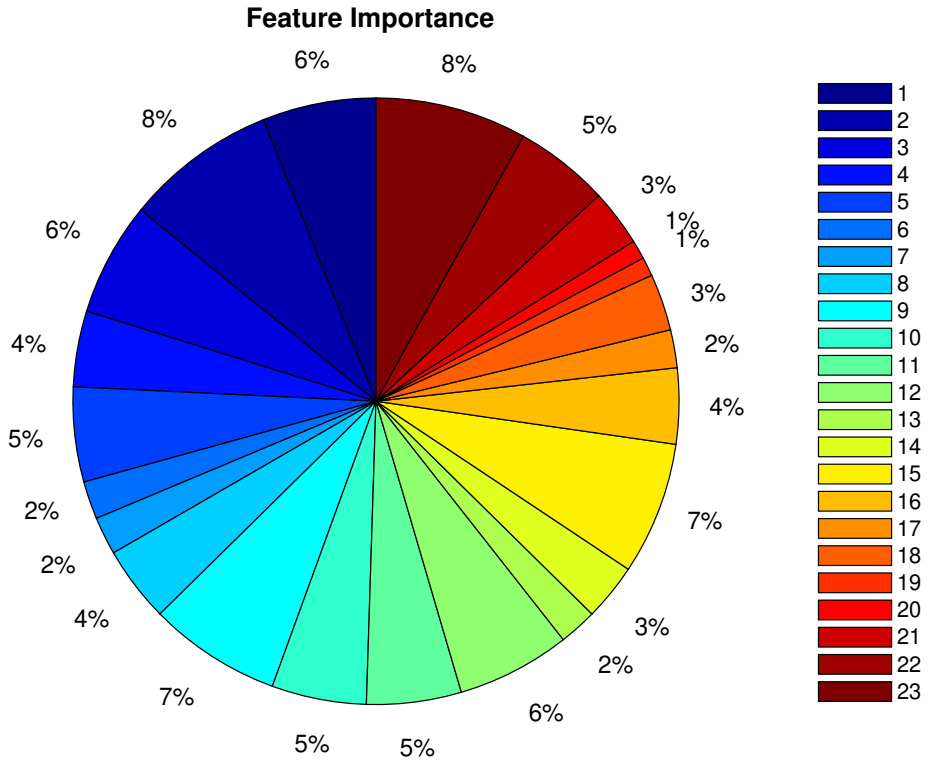

Figure 4.7 – Importance of features based on the overall results of all FS techniques. The feature numbers correspond to the list of features in Table 2.2.

for intrusion detection of the water storage tank seem to be control scheme and pump, which are selected only one percent of the times. The information obtained from the feature selection algorithms can be used to explain the nature of the attack, which in turn helps to plan a suitable response or counter-attack. For instance, one of the top features in Fig. 4.7, namely response address (numbers 2) can be used to detect the reconnaissance attack, as the mismatch between response device addresses is usually an indicator of this attack. Another example is another top feature, time (number 23 in Fig. 4.7), which can be used to detect three types of cyber-attacks such as malicious command injection, malicious response injection, and DOS attacks. The time interval between packets is almost consistent during the normal operation;

however, this measurement becomes very different when such attacks exist in the network. Therefore, one can explain the nature of attack of detecting anomalies in any of these features, as they are indicator of certain events. Knowing the most important features, on the other hand, can inform us regarding the most targeted parts of the system, and its mechanism, which is useful for planning and taking defensive actions.

4.4 Summary

In this chapter, twelve feature selection techniques are reviewed and analyzed on a cyber-physical case study. These case study resembles a SCADA system implemented for a water storage tank, which is under cyber-attacks. The selected feature selection techniques are employed within a multi-modular IDSs, which combines a set of feature selection techniques with two classifiers. This framework enables a comparative study on the feature selection methods, as well as their compatibility with the selected classifiers. Moreover, a feature analysis is performed w.r.t. the results of the feature selection that determines the most important features that are crucial for the task of intrusion detection in the given SCADA system. The feature selection methods in this study achieved satisfying results in terms of accuracy and F1-score. The results indicate that feature selection could improve some certain level of classification accuracy for classifying cyber-attacks. The performed comparative experiment suggests the best combination of feature selection algorithm with the classifiers, and suggests which features should be included for classifying cyber-attacks.

Chapter 5

Conclusions and Remarks

Dimensionality reduction and feature selection are one of the requirement of IDS systems due to the fact that they can directly affect the performance of the detection mechanism. As previously discussed, there are several limitations with current IDS, so in this thesis the promising solutions to overcome the curse of dimensionality as well as low accuracy are proposed, which can enhance the overall IDS performance. These include implementing feature selection and dimensionality reduction techniques along with the classifier in order to enhance the detection accuracy. In this section we compare the results attained through feature selection and diensionality reduction and show which methods perform best in our system.

5.1 Conclusion and Discussion

As discussed earlier, the requirement of feature selection and dimensionality reduction techniques are essential since they provide the most informative features and obtain a subset of network traffic attributes that has no irrelevant and correlated features. Also, the redundant features are reduced in the selected subset. Feature selection and dimensionality reduction improve the accuracy as well as the classification performance of the adopted method. FS and DR enable us to decrease the noise from the data and select the most valuable features to be applied during the training session, this step helps the system to avoid fitting the noise. Thus, feature selection and dimensionality reduction enhance the effeciency of intrusion detection system. For instance, in our SCADA dataset without applying FS and DR the average accuracy

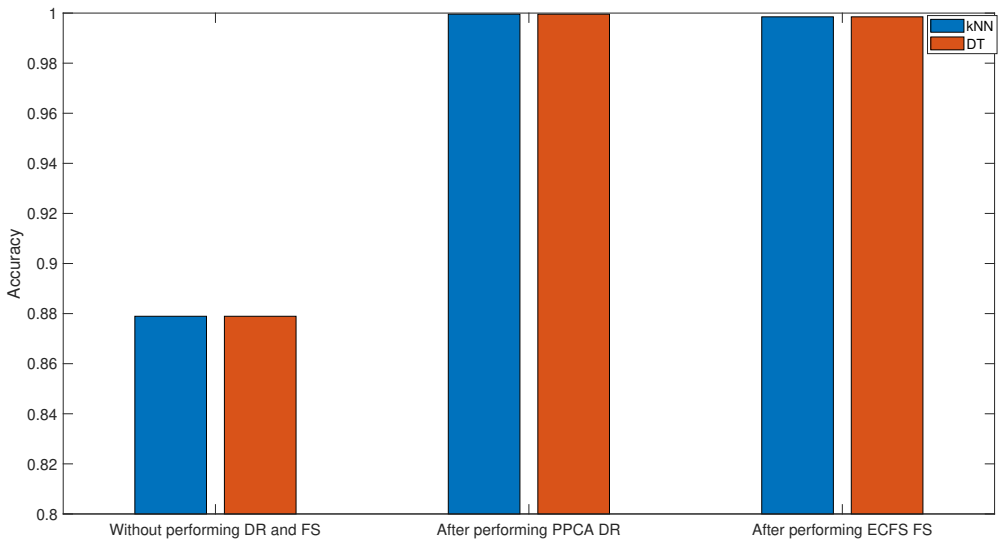

Figure 5.1 – Averaged accuracy obtained by each classifier without performing FS and DR, after performing PPCA, and after performing ECFS.

is around 87% as shown in figure 5.1.

Figure 5.2 compares FS and DR methods with highest performance measures obtained in terms of accuracy and F1-score measures. As shown in the figure, dimensionality reduction methods outperform those obtained through FS methods in terms of accuracy and F1-score. As observed from Fig 5.2, PPCA results in the highest F1-score measure when coupled with kNN 99.97% and ranked second when coupled with DT 99.94%. Then, ECFS method is ranked third when coupled with kNN 99.8% and finally InfFS method along with DT is ranked fourth (99.7%). To wrap up, DR methods proved to be the most effective in removing uninteresting features and hold most of the information in order to make a better detection decision.

Fig 5.3 compares the least dimensionality reduction and feature selection methods observed from our system. LPP method scores 65.8% when coupled with the DT classifier, which results in the lowest accuracy performance among other competitors. PCA method results in a higher performance than LPP with a score 72.6%, when integrated with the kNN classifier. Talking about FS methods, Laplacian scores 82.5%

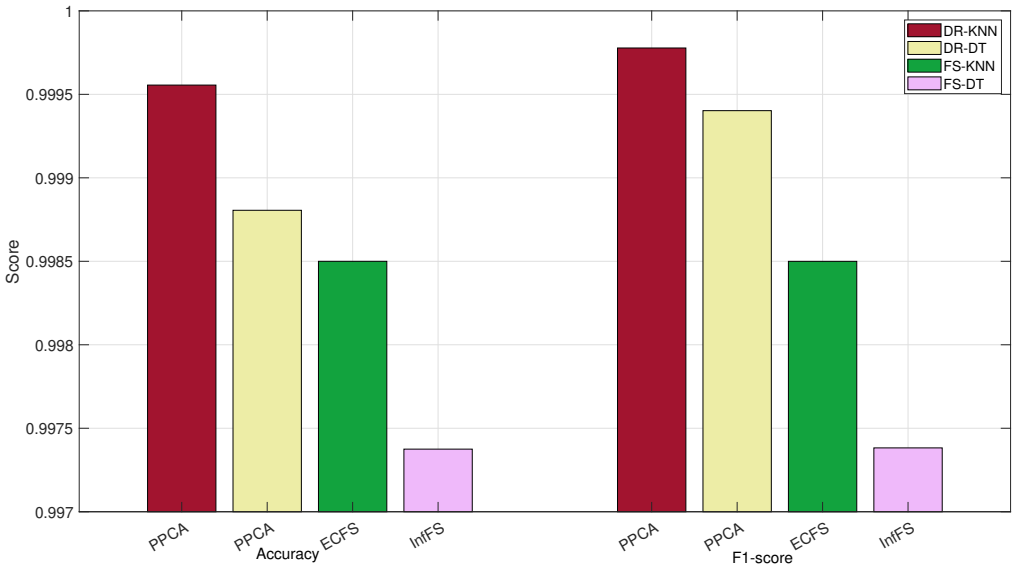

Figure 5.2 – Comparison between top methods of DR and FS.

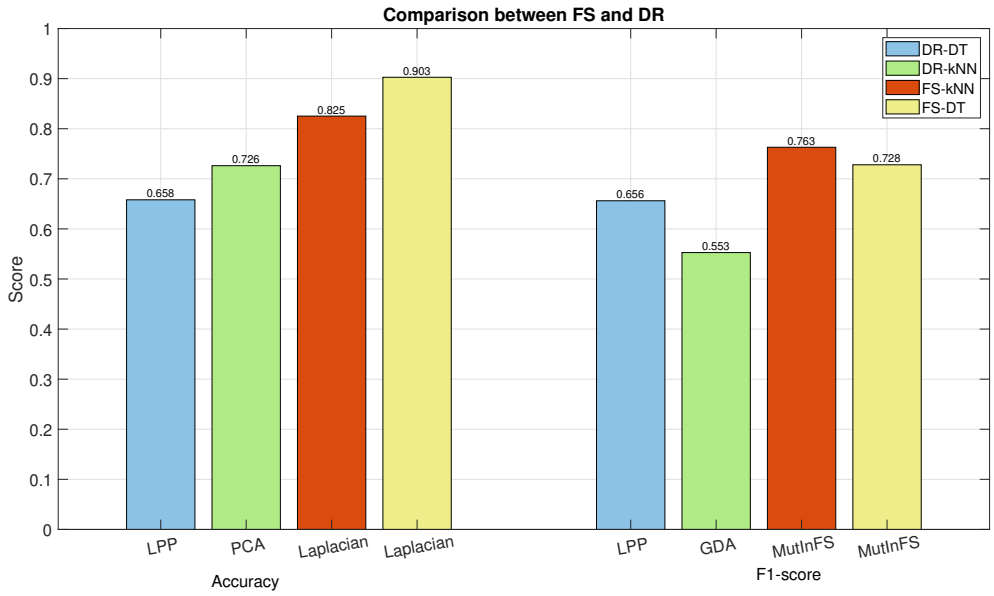

Figure 5.3 – Comparison between least methods of DR and FS.

and 90.3% when combined with kNN and DT that are better than PCA and LPP. Comparing the outcomes in terms of F1-score measure, GDA results in the lowest performance when coupled with the kNN classifer 55.2%. Then, LPP results in the second least performance when combined with the DT classifier 65.6%. In general, MutInFS results in a better performance compared to those combined with the DR methods, when combined with kNN (76.3%) and DT (72.8%).

Although it has been proved that the reduction in number of features can reduce the performance since less number of features provide less amount of information about the network traffic behavior. However, if we select the right number of features, not only the intrusion detection become more efficient in terms of processing time but also the detection accuracy can remain alike to using all the attributes. Generally, PPCA outperforms all other methods, and ECFS method results in the best method among other FS methods in terms of accuracy and F1-score measures, as described previously in Fig 5.2. LPP method attains the lowest accuracy score and GDA results in the lowest F1-score measure. Laplacian results in the lowest accuracy score and MutInFS leads to the lowest F1-score measure among FS competitors.

5.2 Summary

In this thesis, we proposed a data-driven intrusion detection system based on dimensionality reduction and feature selection. We reviewed and explained the basic concepts of different feature selection and dimensionality reduction methods. We examined various predictive models that are built based on different classification techniques such as the k-nearest neighbors and decision tree. We took a brief review of the evaluation criteria that are used to evaluate the predictive models, and rank features that can help to a build stable and robust intrusion detection system.

Our goal was to improve the performance of the intrusion detection system to deal with high dimensional data collected from cyber-physical systems, and to increase the

accuracy and efficiency of the intrusion detection process by eliminating redundant and irrelevant features. We compared a large number of well-known feature selection and dimensionality reduction techniques that are combined with different classifiers for intrusion detection, and, then, the performance of the combined methods is measured in terms of different metrics such as accuracy and F1-score.

As a future research direction, we will verify the proposed intrusion detection scheme by considering a variety of high-dimensional data streams, since the rapid growth of cyber-physical systems poses a challenge to deal with mining these types of data in different contexts.

References

[1] Z. Hammami, M. Sayed Mouchaweh, W. Mouelhi, and L. Ben Said, "Discussion and review of the use of neural networks to improve the flexibility of smart grids in presence of distributed renewable ressources," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1304–1309.

[2] Z. Hammami, M. S. Mouchaweh, W. Mouelhi, and L. B. Said, "Neural networks for online learning of non-stationary data streams: a review and application for smart grids flexibility improvement," *Artif Intelligence Review*, vol. 53, p. 6111–6154, 2020.

[3] R. Razavi-Far, M. Farajzadeh-Zanjani, M. Saif, and S. Chakrabarti, "Correlation clustering imputation for diagnosing attacks and faults with missing power grid data," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1453–1464, 2020.

[4] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16 – 24, 2013.

[5] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[6] H. Hassani, E. Hallaji, R. Razavi-Far, and M. Saif, "A comparative assessment of dimensionality reduction techniques for diagnosing faults in smart grids," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 3618–3623.

[7] E. Hallaji, R. Razavi-Far, and M. Saif, *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*. Research Publishing Services, 2020, p. 1346–1351.

[8] M. Farajzadeh-Zanjani, R. Razavi-Far, and M. Saif, "Dimensionality reduction-based diagnosis of bearing defects in induction motors," in *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017, Banff, AB, Canada, October 5-8, 2017*. IEEE, 2017, pp. 2539–2544. [Online]. Available: https://doi.org/10.1109/SMC.2017.8123006

[9] M. Farajzadeh-Zanjani, E. Hallaji, R. Razavi-Far, and M. Saif, "Generative adversarial dimensionality reduction for diagnosing faults and attacks in cyber-physical systems," *Neurocomputing*, vol. 440, pp. 101–110, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221001521

[10] R. Razavi-Far, E. Hallaji, M. Farajzadeh-Zanjani, M. Saif, S. H. Kia, H. Henao, and G. Capolino, "Information fusion and semi-supervised deep learning scheme for diagnosing gear faults in induction machine systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 8, pp. 6331–6342, 2019.

[11] M. Farajzadeh-Zanjani, R. Razavi-Far, and M. Saif, "A critical study on the importance of feature extraction and selection for diagnosing bearing defects," in *IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018, pp. 803–808.

[12] H. Hassani, E. Hallaji, R. Razavi-Far, and M. Saif, "Unsupervised concrete feature selection based on mutual information for diagnosing faults and cyber-attacks in power systems," *Engineering Applications of Artificial Intelligence*, vol. 100, p. 104150, 2021.

[13] R. Razavi-Far, E. Hallaji, M. Saif, and G. Ditzler, "A novelty detector and extreme verification latency model for nonstationary environments," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 561–570, 2019.

[14] R. Razavi-Far, V. Palade, and E. Zio, *Optimal Detection of New Classes of Faults by an Invasive Weed Optimization Method*. IEEE, Sep. 2014, vol. Article number 6889887, pp. 91–98.

[15] R. Razavi-Far, E. Hallaji, M. Farajzadeh-Zanjani, and M. Saif, "A semi-supervised diagnostic framework based on the surface estimation of faulty distributions," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1277–1286, 2019.

[16] R. Razavi-Far, E. Hallaji, M. Saif, and L. Rueda, "A hybrid scheme for fault diagnosis with partially labeled sets of observations," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 61–67.

[17] M. Farajzadeh-Zanjani, E. Hallaji, R. Razavi-Far, M. Saif, and M. Parvania, "Adversarial semi-supervised learning for diagnosing faults and attacks in power grids," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3468–3478, 2021.

[18] E. Hallaji, R. Razavi-Far, and M. Saif, "Detection of malicious SCADA communications via multi-subspace feature selection," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[19] M. Sayed-Mouchaweh, Ed., *Diagnosability, Security and Safety of Hybrid Dynamic and Cyber-Physical Systems*, 1st ed. Springer International Publishing, 2018.

[20] V. M. Igure, S. A. Laughter, and R. D. Williams, "Security issues in scada networks," *Computers & Security*, vol. 25, no. 7, pp. 498 – 506, 2006.

[21] geeksforgeeks, "Intrusion detection system," in *IDS*, 2020. [Online]. Available: https://www.geeksforgeeks.org/intrusion-detection-system-ids/

[22] A. Alazab, M. Hobbs, J. Abawajy, and M. Alazab, "Using feature selection for intrusion detection system," in *2012 international symposium on communications and information technologies (ISCIT)*. IEEE, 2012, pp. 296–301.

[23] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1184–1199, 2011.

[24] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur, "Bayesian event classification for intrusion detection," in *19th Annual Computer Security Applications Conference, 2003. Proceedings.* IEEE, 2003, pp. 14–23.

[25] L. Koc, T. A. Mazzuchi, and S. Sarkani, "A network intrusion detection system based on a hidden naïve bayes multiclass classifier," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 492–13 500, 2012.

[26] H. Shapoorifard and P. Shamsinejad, "Intrusion detection using a novel hybrid method incorporating an improved knn," *Int. J. Comput. Appl*, vol. 173, no. 1, pp. 5–9, 2017.

[27] S. Vishwakarma, V. Sharma, and A. Tiwari, "An intrusion detection system using knn-aco algorithm," *Int J Comput Appl*, vol. 171, no. 10, pp. 18–23, 2017.

[28] R. Razavi-Far, M. Farajzadeh-Zanjani, S. Chakrabarti, and M. Saif, "Data-driven prognostic techniques for estimation of the remaining useful life of lithium-ion batteries," in *IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2016, pp. 1–8.

[29] M. Farajzadeh-Zanjani, R. Razavi-Far, and M. Saif, "Efficient sampling techniques for ensemble learning and diagnosing bearing defects under class imbal-

anced condition," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–7.

[30] S. Chakrabarti, R. Razavi-Far, M. Saif, and L. Rueda, "Multi-class heteroscedastic linear dimensionality reduction scheme for diagnosing process faults," in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2017, pp. 1–4.

[31] T. Morris, A. Srivastava, B. Reaves, W. Gao, K. Pavurapu, and R. Reddi, "A control system testbed to validate critical infrastructure protection concepts," *International Journal of Critical Infrastructure Protection*, vol. 4, no. 2, pp. 88–103, 2011.

[32] T. Morris and W. Gao, "Industrial control system traffic data sets for intrusion detection research," in *International Conference on Critical Infrastructure Protection*. Springer, 2014, pp. 65–78.

[33] R. Razavi-Far, S. Chakrabarti, M. Saif, and E. Zio, "An integrated imputation-prediction scheme for prognostics of battery data with missing observations," *Expert Systems with Applications*, vol. 115, pp. 709 – 723, 2019.

[34] R. Razavi-Far, M. Farajzadeh-Zanjani, B. Wang, M. Saif, and S. Chakrabarti, "Imputation-based ensemble techniques for class imbalance learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 05, pp. 1988–2001, 2021.

[35] V. Sumithra and S. Surendran, "A review of various linear and non linear dimensionality reduction techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, pp. 2354–2360, 2015.

[36] A. Navlani, "Introduction to factor analysis in python," *data camp*, 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/introduction-factor-analysis

[37] S. Solutions, "Confirmatory factor analysis," *Retrieved May*, vol. 28, p. 2016, 2013.

[38] J. Imperial, "The multidimensional scaling (mds) algorithm for dimensionality reduction," *medium- Data Driven Investor*, Aug. 2019.

[39] S. Raschka, "Linear discriminant analysis," *sebastianraschka*, Aug. 2014.

[40] G. Rosman, M. M. Bronstein, A. M. Bronstein, and R. Kimmel, "Nonlinear dimensionality reduction by topologically constrained isometric embedding," *International Journal of Computer Vision*, vol. 89, no. 1, pp. 56–68, 2010.

[41] P. Bafna, S. Shirwaikar, and D. Pramod, "Task recommender system using semantic clustering to identify the right personnel," *VINE Journal of Information and Knowledge Management Systems*, 2019.

[42] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[43] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[44] A. K. PAL, "Dimension reduction - isomap," *paperspace*, Apr. 2018.

[45] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[46] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM Journal of Scientific Computing*, vol. 26, pp. 313–338, 2002.

[47] K. C. Kempfert, Y. Wang, C. Chen, and S. W. Wong, "A comparison study on nonlinear dimension reduction methods with kernel variations: Visualization, optimization and classification," *Intelligent Data Analysis*, vol. 24, no. 2, pp. 267–290, 2020.

[48] F. Bahmaninezhad and J. H. Hansen, "Generalized discriminant analysis (gda) for improved i-vector based speaker recognition." in *Interspeech*, vol. 2016, 2016, pp. 3643–3647.

[49] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1208–1213.

[50] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[51] J. De la Porte, B. Herbst, W. Hereman, and S. Van Der Walt, "An introduction to diffusion maps," in *Proceedings of the 19th Symposium of the Pattern Recognition Association of South Africa (PRASA 2008), Cape Town, South Africa*, 2008, pp. 15–25.

[52] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[53] S. Sun and Q. Chen, "Hierarchical distance metric learning for large margin nearest neighbor classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 07, pp. 1073–1087, 2011.

[54] R. Razavi-Far and M. Kinnaert, "Incremental design of a decision system for residual evaluation: A wind turbine application," *IFAC Proceedings Volumes*, vol. 45, no. 20, pp. 343–348, 2012.

[55] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4202–4210.

[56] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, "Infinite latent feature selection: A probabilistic latent graph-based ranking approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1398–1406.

[57] B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary computation algorithms for feature selection of eeg-based emotion recognition using mobile sensors," *Expert Systems with Applications*, vol. 93, pp. 143–155, 2018.

[58] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.

[59] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, "Feature selection via mutual information: New theoretical insights," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.

[60] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," *arXiv preprint arXiv:1908.05376*, 2019.

[61] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines." in *ICML*, vol. 98, 1998, pp. 82–90.

[62] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2006, pp. 507–514.

[63] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 333–342.

[64] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83–90, 2006.

[65] J. Han, Z. Sun, and H. Hao, "l0-norm based structural sparse least square regression for feature selection," *Pattern Recognition*, vol. 48, no. 12, pp. 3927–3940, 2015.

[66] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.

Vita Auctoris

Ranim Aljoudi was born in 1996 in Syria. Studied her High-school in Dubai then pursued her undergraduate studies in the school of Computer Science at University of Windsor, and received B.Sc. degree Honors in Applied Computing in 2017 . She is currently a candidate for the MASc degree in Electrical and Computer Engineering at the University of Windsor, Canada and expects to graduate in Spring 2021. Her research area mainly involves machine learning, cyber security, data mining, and their applications.