A Thesis

entitled

Distributed Intrusion Detection System in A Multi-Layer Network Architecture of Smart Grids

by

Yichi Zhang

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Electrical Engineering

_____
Dr. Lingfeng Wang, Advisor

_____
Dr. Weiqing Sun, Co-Advisor

_____
Dr. Mansoor Alam, Committee Member

_____
Dr. Patricia R. Komuniecki, Dean
College of Graduate Studies

The University of Toledo

August 2011

An Abstract of

Distributed Intrusion Detection System in A Multi-Layer Network Architecture of Smart Grids

By

Yichi Zhang

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Electrical Engineering

The University of Toledo
August 2011

This thesis proposes a Distributed Intrusion Detection System for Smart Grids by developing and deploying intelligent modules in multiple layers of the smart grid in order to handle cyber security threats. Multiple Analyzing Modules are embedded at different levels of the smart grid - the Home Area Network, Neighborhood Area Network, and Wide Area Network. These intelligent modules employ Support Vector Machines and Artificial Immune System to detect and classify malicious data and possible cyber attacks. Analyzing Modules at different levels are trained using data that are relevant to their levels and will also be able to communicate with each other in order to improve the detection performance. Simulation results demonstrate that this is a promising methodology for improving system security through the identification of malicious network traffic, and the detection efficiency is improved by applying the optimal communication routing.

For my parents and friends

Acknowledgements

I would like to express my deepest appreciation to my advisor Dr. Lingfeng Wang and co-advisor Dr. Weiqing Sun. They have provided me with quality instructions, continuous guidance, and unconditional support throughout my study at the University of Toledo. They gave me the freedom to explore new ideas and always offered me constructive advices at the right times. Also, I wish to thank Robert C. Green II, and all the other Lab members for their support and feedback on various parts of this thesis. I am grateful for having been among you during my Master's study here in Toledo. Finally, I would like to thank my family and all my friends. Thank you for giving me the support and love to complete this work.

Table of Contents

Abstract iii

Acknowledgements v

Table of Contents vi

List of Tables viii

List of Figures ix

List of Abbreviations xi

1 Introduction 1

1.1 Background 1

1.2 Research Approach 7

1.3 Organization 8

2 Literature Review 9

2.1 Smart Grid 9

2.2 Cyber Security Issues in Smart Grid 13

2.3 Intrusion Detection System Background 17

2.4 Computational Intelligence Algorithms in the Intrusion Detection Systems 20

2.5 Intrusion Detection in Smart Grid 23

3 System Design 27

3.1 Network Architecture 27

3.2 Distributed IDS Modules 31

3.3 Wireless Mesh Network Model and the Optimal Routing Algorithm 33

3.4 SVM for Attack Classification 36

3.5 Clonal Selection Classification Algorithm for Attack Detection 38

3.6 Dataset and Preprocessing 43

4 Simulations Results and Analysis 46

4.1 Distribution of Possible Attacks in Different Layers of SGDIDS 46

4.2 Grid Topology and Communication Routing Optimization 47

4.3 Intrusion Detection with the SGDIDS 51

4.4 Results and Discussion 56

5 Conclusions and Future Work 60

5.1 Conclusions 60

5.2 Future Work 62

References 64

Appendices

A Screenshots of Weka Platform for AIS Classifiers 76

A.1 Preprocess Window (User-Defined) 76

A.2 Classifier Selection Window (User-Defined) 77

A.3 Classification Using CLONALG Model 78

A.4 Classification Using AIRSParallel Model 79

B Source Code for Primal-Dual Optimal Algorithm and Classifier Algorithms (Matlab) 80

B.1 Source Code for Primal-Dual Routing Optimization Algorithm 80

B.2 Source Code for AIS Classifier Selection Algorithm 82

List of Tables

4.1 The Routing Table of the Communication Among Nodes 51

4.2 The Influencing Parameters and the Optimized Values 52

4.3 The Routing Table of the Communication among Nodes 56

4.4 The Overall FPR/FNR Values of the DIDS 58

List of Figures

1.1 Interaction between Institutes in the Smart Grid through Secure Communication and Electricity Flows. 4

1.2 The CIA-triad. 5

1.3 The Interconnection Core between the Smart Grid and Other Systems. 6

2.1 The Structure of RBFNN at the Detecting Stage. 21

2.2 The Agent Using GP. 22

3.1 Three Layer Network Architecture. 29

3.2 Home Area Network Intrusion Detection System. 32

3.3 Analyzing Module Architecture. 32

3.4 Neighborhood Area Network Intrusion Detection System. 33

3.5 Wide Area Network Intrusion Detection System. 33

3.6 Primal Dual Algorithm. 35

3.7 Diagram of the CLONALG algorithm. 41

3.8 The life cyber of the AIRS algorithm. 42

4.1 Smart Grid Distributed Intrusion Detection System Topology. 49

4.2 Optimal Communication Routing Selection in the SGDIDS. 50

4.3 Intrusion Detection Routing in Home, Neighborhood, and Wide Area Network. 53

4.4 Detection Procedure in the SGDIDS. 55

4.5 Results Using SVM and AIS Algorithms..... 58

List of Abbreviations

AMI	Advanced Metering Infrastructure
ACO	Ant Colony Optimization
ADF	Automatically Defined Function
AIRS	Artificial Immune Recognition System
AIS	Artificial Immune System
AM	Analyzing Module
AMI	Advanced Metering Infrastructure
ANN	Artificial Neural Network
CAC	Central Access Controller
CIDS	Centralized Intrusion Detection System
CIP	Common Industrial Protocol
CM	Controlling Module
CSA	Clonal Selection Algorithm
CSCA	Clonal Selection Classification Algorithm
DB	Data Bus
DSM	Data Segmentation Module
E&S C	Energy and Service Corporations
EC	Evolutionary Computation
EDS	Energy Distribution System
GA	Genetic Algorithms
GP	Genetic Programming
HAN	Home Area Network
IAM	Information Acquisition Module
KDD99	KDD Cup 1999 Dataset
MIM	Man-in-the-middle
MM	Metering Module
NAN	Neighborhood Area Networks

NASPInet	North American SynchroPhasor Initiative network
NN	Neural Network
OM	Output Module
PLC	Power Line Communication
PM	Preprocessing Module
PMU	Phasor Measurement Unit
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RTU	Remote Terminal Unit
SAL	System Abstraction Layer
SCADA C	Supervisory Control And Data Acquisition Controller
SGDIDS	Distributed Intrusion Detection System for Smart Grids
SM	Service Module
SMDC	Smart Meter Data Collector
SOM	Self-Organizing Maps
SoS	System of Systems
SVM	Support Vector Machine
WAMS	Wide Area Measurement System
WAN	Wide Area Network
WMN	Wireless Mesh Network

Chapter 1

Introduction

1.1 Background

It is predicted that the amount of the electricity utility will increase 30 percent in the upcoming 25 years. The electricity grid we are using in current days, however, was built in the early 20th century and could not satisfy such great amounts of the electricity. The power grid is becoming increasingly overloaded and unstable. Some of the infrastructures have been obsolete to transmit the electricity, and the blackouts may have more chances of occurrences [1], [2].

Since 1982, growth in peak demand for electricity driven by population growth and various digital products, such as high-definition televisions, computers and video game systems, more air conditioners and more computers, has exceeded transmission growth by almost 25% every year. From 1998, the frequency and magnitude of outages have increased at an alarming rate. There have been five massive blackouts over the past forty years, and three of them occurred within the nine years. And according to the data record, billions of dollars would be wasted every minute as the economic loss in the blackouts in U.S. For instance, one of recent blackouts happened on November 11, 2009 in Brazil [3]. It lasted 4 hours and nearly half of the country was plunged into darkness. And life of more than one

third of citizens was affected. It was regarded as "world's worst power cuts". What's more, varieties of uncertainties are created by growing concerns over power plant pollution. In order to prevent the disasters brought by the blackouts, an affordable, reliable, secure, efficient, environmentally friendly grid is needed to meet the needs of the customers.

The concept of the smart grid promises the world an efficient and intelligent approach of managing energy supply and consumption. Consumers and energy suppliers alike can take advantage of the convenience, reliability, and energy savings provided through real time energy management. The "smart gird" is composed of the sensors, digital smart meters, digital controls, and analytics tools to monitor and control two-way energy flow [4]. This enables the customers to manage their energy based on the information from their energy consumption and management tools in the individual networked appliance. The customers could apply the optimal algorithms to purchase the cheapest electricity depending on the amount of electricity consumptions in different time; also the customers can generate their own electricity and sell it back to the grid. Some sensors will also be able to monitor the damaging situation of the smart grid, the power outages could be detected immediately, and the damages will be located and isolated for failure reparation. Also, the customers will be informed when the power is restored on time. What's more, the smart grid will allow the energy suppliers and utility companies to aware power demands from the users in real time, so that the delivery and incorporation of the energy could be improved. These capabilities enable the application of the renewable energy like wind and solar power, and will achieve the rapidly increasing energy needs around the world such as the application of power efficient devices like plug-in electric vehicles [5], [6].

One of the advantages of the smart grid is the installation of a completely new, two-way communication network between energy suppliers and their customers. This allows the

smart grid to be viewed as an electric power grid that has an integrated data communication network allowing the collection and analysis of data by the communication sensors at all levels in real time [7]. The real-time communication ability of the smart grid will enable utilities to optimize and modernize the power grid in order to realize its full potential [8]. The requirement of the communications for the data transport of the smart grid is to address the backbone and the spur segments. The smart grid will use the upgraded backbones as backhaul Ethernet/IP data traffic with the highest speed at 10 GB/S to achieve in the highly reliable manner. The existing networks could be overlaid by overbuilding the gigabit Ethernets on the available wired or broadband wireless network over the microwave paths [9]. The communication network of the smart grid will provide a number of new energy concepts including real-time pricing, load shedding, consumption management, cost savings from peak load reduction and energy efficiency, integration of plug-in hybrid electric vehicles for grid energy storage, and the integration of alternative distributed generation sources including photovoltaic systems and wind turbines. This new communication network will be constructed using various communication paths including fiber optic cable, twisted pair, broadband over power line, and wireless technologies [10]. By providing a framework identifying seven domains within the smart grid—Transmission, Distribution, Operations, Bulk Generation, Markets, Customer, and Service Provider, a secure communication architecture is provided by [11]. In it, a smart grid domain is defined as a high-level grouping of organizations, buildings, individuals, systems, devices, or other institutes with similar objectives and similar types of applications. Across the seven domains, numerous institutes will capture, transmit, store, edit, and process the information necessary for smart grid applications. To enable the functionality of the smart grid, the institutes in the

particular domain can interact with institutes in other domains. The secure communication architecture proposed from [11] and [12] is illustrated in Figure1.1.

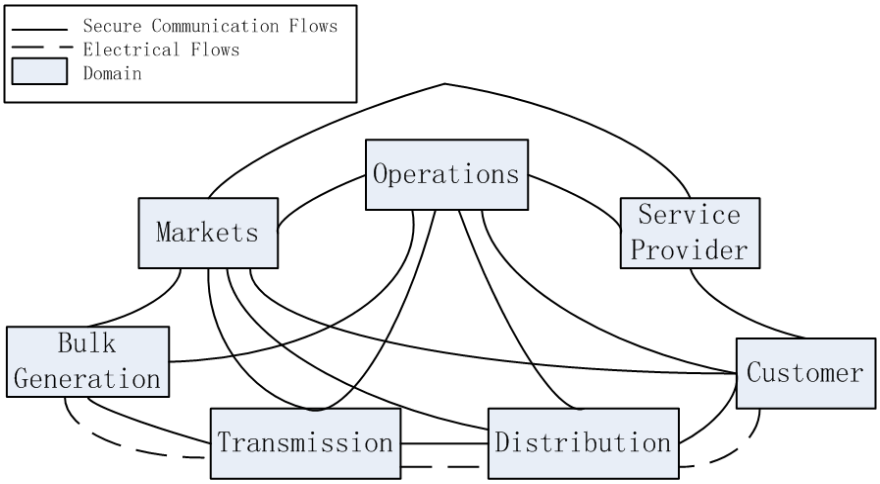

Figure 1.1: Interaction between Institutes in the Smart Grid through Secure Communication and Electricity Flows [11], [12]

The proliferation of these new technologies, especially an Internet-like communications network, may introduce some new threats to the security of the smart grid. In the smart grid network there are three crucial aspects of security that may be threatened due to the CIA-triad. The CIA-triad defines three principles of security issues: Confidentiality, Integrity, and Availability [13]. These principles are illustrated in a triangle secure system in Figure1.2. The three aspects of the CIA-triad are defined as follows.

Confidentiality is "the property that information is not made available or disclosed to unauthorized individuals, entities or processes". The attack of the confidentiality occurs when an unauthorized person, entity or process can reach into the system and access the information.

Integrity is "the property of safeguarding the accuracy and completeness of assets". The integrity of the system proves the information in the system won't be modified by the attacks.

Availability refers to "the property of being accessible and usable upon demand by an authorized entity". The resources should be kept accessible at all times to authorized entities or processes [14].

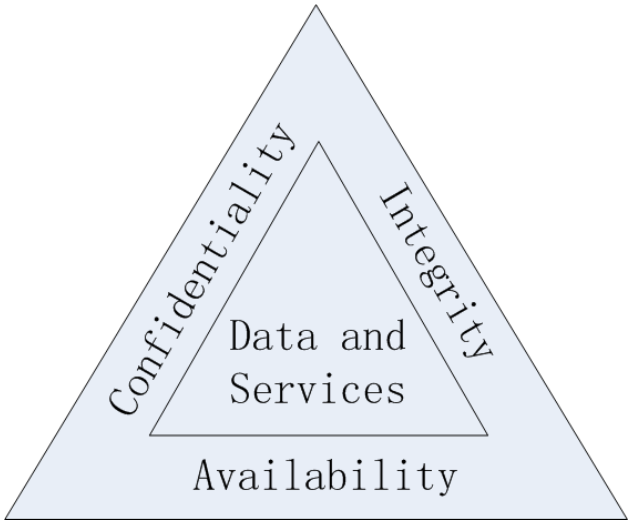

Figure 1.2: The CIA-triad [8]

Besides the application of the traditional computer network security mechanisms such as secure protocols, intrusion detection systems, and appropriate security processes, the smart grid should also considers novel challenges and attacks for its cyber security. The computer network security approaches should be applied in the industrial control systems; also it should take into account the real-time nature of the smart grid and adapt the risk management as graceful degradation when the grid is attacked. The interconnection between the smart meters or other smart grid components and the systems, such as the building networks, may also bring cyber security challenges on trust, key management, and the relative authorization policies [15]. The interconnection between the smart grid and the other systems is illustrated in Figure 1.3.

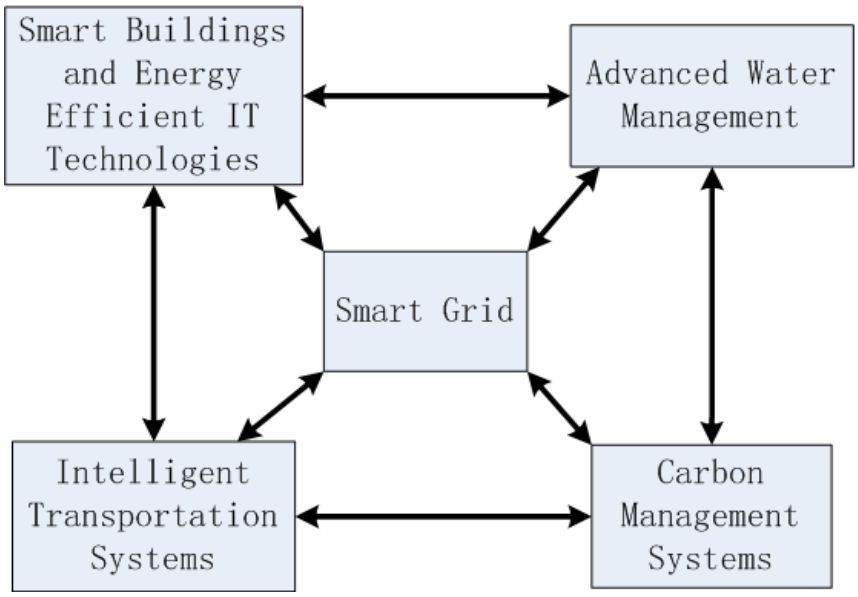

Figure 1.3: The Interconnection Core between the Smart Grid and Other Systems [15]

Beyond cyber threats like malware, spyware, and computer viruses that currently threaten the security of computer communication networks, the introduction of new and distributed technologies such as smart meters, sensors, and other sub networks can bring new vulnerabilities to the smart grid [16]. For instance, the data flow of the entire network may be interrupted using SYN flood attacks against the metering infrastructure. In the smart metering systems, several ports are applied for transmitting metering and controlling information from the customers to the energy company. These new institutes will bring new attacks to the grid system. For instance, P0 is the port which provides administrative access to service engineers, and the customers can access and try to connect to their metering systems through the right equipment to interface with P0. Due to the configuration options available through P0, the meter data or price settings can be altered and the data integrity could be breached. Also, availability of the meter data will be affected when communication settings with other connected ports are modified. Another instance of P0 security is when an engineer readout or modify meter settings through a set of configuration buttons on the

meter combined with an optional display. In case poor protective measures such as a secret key combination are being applied to access these administrative features, this may lead to perform attacks without using any additional hardware [8].

Another aspect is the electronic security of critical cyber assets of the power infrastructure in the power community. It includes the supervisory control and data acquisition (SCADA) systems that are widely applied in the industry for monitoring and control of the power grid. The SCADA systems include computer and communication devices installed in power plants, substations, energy control centers, company headquarters, regional operating offices, and large load sites. In the three main control systems of the critical infrastructure systems, the SCADA systems are the central nerve system of a wide-area control network that constantly gathers the latest status from remote units [17]. The communication system for wide-area protection and control of a power grid can be blocked or cutoff due to component failures or communication delays [18]. If one of the crucial communication channels fails connecting in the operational environment, the inability to control or operate important facilities may happen, and possible power outages may be raised [19].

1.2 Research Approach

Due to the issues presented in 1.1, an integrated strategy to improve the security of the power grid with regards to the cyber-physical security of the smart grid is extremely important. In the following chapters we propose a new architecture for a hierarchical and distributed intrusion detection system called the smart grid Distributed Intrusion Detection System (SGDIDS) that is applied to a representative three-layer communication network. This Distributed Intrusion Detection System (DIDS) is able to successfully analyze

communication traffic using an Analyzing Module (AM) that leverages classification algorithms such as Support Vector Machine (SVM) and Artificial Immune System (AIS) in order to determine if an attack is occurring, what type of attack it is, and where it comes from in the communication system.

1.3 Organization

The remainder of this thesis is organized as follows: Chapter 2 provides a literature review of this research field, Chapter 3 presents the architecture and design of the SGDIDS proposed, Chapter 4 gives the simulation results and analysis of the SGDIDS in multiple scenarios, and Chapter 5 provides conclusions and future work.

Chapter 2

Literature Review

2.1 Smart Grid

The electric transmission and distribution system is on the verge of a transformation that will integrate the advanced communication network capabilities of the information technology sector with the traditional electricity delivery capabilities of the existing transmission and distribution grid, this technologically enhanced electric grid is named as the "smart grid" [1], [20]. This advance in the electric system promises improvements in how the power is transmitted and used, and the new relationships between the commissioners, companies, customers, and other electricity sector stakeholders should be constituted.

The transformation from the current power grid to the smart grid requires new investment and commitment with high value returned. Expectations from the smart grid can be divided into six value areas.

- More reliable: the new smart grid can provide enough and qualified power when the users need it on time.

- More secure: the smart grid withstands physical and cyber attacks and prevent the system from suffering massive blackouts or exorbitant recovery costs. It will be less vulnerable to natural disasters and can recover from the attacks quickly.
- More economic: the more economic smart grid will be operated under the basic laws of supply and demands, and the fairer prices of power could be supplied.
- More efficient: new strategies should be applied to provide the smart grid characteristics of cost control, minimal transmission and distribution losses, efficient power production, and optimal asset utilization, the consumers will also manage their energy usage.
- More environmentally friendly: the smart grid will reduce environmental impacts through improvements in efficiency and applying more intermittent and renewable resources so that the pollution and greenhouse effect would be prevented.
- Much safer: the smart grid will not be harmful to the public or the grid workers and will be sensitive to users who depend on it as the medical utilization.

The smart grid can be considered as the transactive agents which enables financial, informational, and the electrical transactions among the consumers, energy suppliers, and the other authorized users. The smart grid has the following seven principal characteristics [21].

- Enable active participation by consumers: the smart grid will provide consumers information, control, and options so that they could be engaged in the new electricity markets. The willing consumers can modify their consumptions based on the balancing of their demands and resources which are supported by the grid operators, so that their electric equipments can reach higher level of requirements.

- Accommodate all generation and storage options: all types and sizes of electrical generation and storage systems will be integrated by simplified interconnection processes and interoperability standards to achieve a "plug-and-play" level of convenience. Large central power plants will utilize renewable resources to develop and deploy the distributed equipments such as the electric vehicles.
- Develop new products, services, and markets: the smart grid will connect the consumers and the energy suppliers together. This will support the creation of new electricity markets from the home energy management system to the technologies that allow consumers and third parties to bid their energy resources into the electricity market. A consistent market operation across regions will be supplied by the smart grid.
- Provide power quality for the digital economy: the smart grid will monitor, diagnose, and respond to the power quality deficiencies and prevent a dramatic reduction of business losses to the consumers resulted from insufficient power quality.
- Asset utilization optimization and efficiency: the smart grid will improve the load factors, outage management performance, and lower the system losses. It has great additional space to build and repair the necessary equipment by the planners and engineers, and the management to the work force will be more effectively.
- Self-healing ability: the smart grid will have the ability of anticipating and responding to the system disturbances. By applying the continuous self-assessments and detecting the attacks, the smart grid will take corrective actions, restore the grid components or network sections, and rapidly heal itself. The problems of civil alteration and management will also be settled by changeable network.

- Resilient operations for cyber and physical attacks: the smart grid will apply a system based solution to reduce cyber and physical vulnerabilities and recover rapidly from them. The resilience of the smart grid will store a database for intrusion prevention, and it will also be built more robustly to the natural disasters.

The development of the smart grid is seen as four milestones, completion of these milestones requires development and integration of various technologies and applications.

- Consumer Enablement (CE): the CE is the current development procedure of the smart grid, which empowers the consumers by supplying them necessary information of the smart grid for the new equipment and options effectively. It includes information of Advanced Metering Infrastructure (AMI), home area networks (HAN) with in-home displays, distributed energy resources (DER), demand response programs, and upgrades to utility information technology architecture and applications with all future technologies.
- Advanced Distribution Operations (ADO): ADO will improve reliability of the communication and energy supply, and will enable the self-healing of the network. ADO includes equipments such as smart sensors and control devices, advanced outage management, distribution management and distribution automation systems, geographical information and other technologies which are used for 2-way power flow and micro-grid operation.
- Advanced Transmission Operations (ATO): ATO combines the distribution system with Regional Transmission Organization operational and market applications, improves overall grid operations, and will reduce transmission congestion. ATO includes substation automation, integrated wide area measurement applications, power electronics, advanced system monitoring and protection schemes and

modeling, simulation, and visualization tools to increase situational awareness and supply the understanding of real time and operating risks in the future.

- Advanced Asset Management (AAM): AAM will integrate the grid intelligence acquired after completing the above three milestones with new and existing asset management applications. This integration will enable reducing the operations and maintenance costs of the equipments and achieve better utilizing assets during operations. Also, AAM will significantly improve the performance of capacity planning, maintenance, engineering and facility design, customer service processes, and work and resource management [22].

2.2 Cyber Security Issues in Smart Grid

The cyber security of the smart grid is quickly becoming an important issue due to the growing awareness of security issues in cyberspace. The potential vulnerabilities and attacks brought about by the inclusion of new technologies that will support the smart grid are discussed in [8], [17], [24], [25]. In [26] it is found that the confidentiality or integrity of the consumers' data may be compromised due to the increase of the entry points and paths for adversaries, or the introduction of malicious software. Therefore, the tasks of the cyber security of the smart grid are described in [27] such as selecting the use cases with cyber security considerations and performance of a risk assessment.

The current specific issues of cyber security in the smart grid applications focus mostly on topics such as [23]:

- AMI Security
- Privacy of customer data
- Protocol for the mapping of different players in different domains

- Risk Management for SCADA systems
- Cyber Security Risk Management Framework

Several parts of the smart grid may suffer from cyber security vulnerabilities such as the Advanced Metering Infrastructure (AMI), unsecured wireless mesh network communications, and other attackable technologies.

The AMI is composed of four parts (A smart meter, customer gateway, AMI communication network, and headend) and can generally be considered as a metering system providing two-way communications, automated meter data collection, outage management, dynamic rate structures, and demand response for load control [13]. It is viewed as a fundamental technology for the smart grid. Currently, there are a number of potential vulnerabilities that have been discovered in AMI networks. For instance, the device memory in the AMI network could be modified by inserting malicious software [28] or the disconnect command may be sent to the meters which would block the transmission of metering information [13]. AMI system security must protect the tasks of all AMI business functions without introducing the attacks as some method of control of the grid. This does not imply that AMI security architects are only responsible for ensuring this, but rather that responsibility must be assigned for a system perspective wherein potential AMI impacts on the larger grid are analyzed, anticipated, and defended against in some portion of the whole system of systems (SoS) architecture and implementation [29].

Wireless networks are also commonly used in the current smart grid because of their convenience and low cost. Some smart grid implementations use mesh networks with wireless devices to provide self-adapting, multi-path, or multi-hop communication between the nodes. Mesh networks provide redundant communication paths as well as multiple communication paths that can compensate for the failures brought on by the break-down of

communication nodes. For instance, Zigbee is a low-cost and low-power wireless mesh networking protocol which can be widely deployed and utilized in a more reliable environment. This makes it advantageous to use the technology in the smart grid even though it allows for easy attacks due to the commoditized and open nature of the wireless radios used. An example of this can be found in [30] where DoS attacks are associated with Zigbee implementations in the smart grid. Further examples of attacks that are used against wireless network technologies include damaging the integrity of configuration, routing, and communication traffic, illegitimate network operations, inconsistent traffic direction, alteration of data, unintentional dissemination of consumer data, unresponsive destination nodes, grid bandwidth over-usage and signal power over-consumption. For the wired networks, various researches on the security are also being developed, for instance, In [31], the security of the wide area measurement system (WAMS) that can be applied in the wide area communication networks of the smart grid is researched and the development of the WAMS technologies are expected to be become integrated into the real-time control system.

Each two-way communication path which supports control and measurement in the smart grid also has the potential to become an entry point for both physical and cyber attacks that may be used by anyone with mal intent. Wireless networks can easily be probed or sniffed by attackers and are susceptible to Man-in-the-Middle (MIM) attacks. Although there are security mechanisms which could prevent unauthorized use of these communication paths, weaknesses still exist in these mechanisms. In [31], it found the cyber security vulnerabilities in the North American SynchroPhasor Initiative network (NASPInet) it proposed, such as intrusions caused by the insufficient fault tolerant design that avoids single points of failure; the non-authorized admission control connected to the Data Bus

(DB); the attacks to end-to-end integrity and confidentiality of the Phasor Measurement Unit (PMU) data; and the malicious logging and auditing to the files.

What is more, in smart meters and SCADA systems, it is possible to log on to these nodes and reprogram the measuring and controlling commands [33]. For instance, in the smart meter systems, one crucial issue is the integrity of the meter reading to the utility. This requires the meter should not be reporting false values under any circumstance. However, this is not realistic because the human interaction always exists and the access control of the smart meter is operated by the consumers [13]. This could lead to significant errors in power measurements which could, in turn, lead to severe power outages. Also, without a fixed standard or proprietary security mechanism, the implementation of the smart grid may lead to poor interoperability and issues of security by obscurity. This means that if the security of the smart grid is designed based on the concept of hiding all of the known vulnerabilities and flaws, knowledgeable attackers may overcome network security and the grid with ease [34]. Although a series of regulations for critical infrastructure protection have been published, a mature standard is still in process.

In the portion of the power system dealing with monitoring and control, a number of digital vulnerabilities have been discovered at the entrances to substations of the SCADA systems [35]. Multiple access points of the SCADA could be attacked. Examples of this include blocking or probing the communications between the modem and the Remote Terminal Unit (RTU). A DoS attack is simulated in [36] where clients communicate with the simulation of a power system that was implemented using PowerWorld. In [37], it simulated a shorted power line in a SCADA power grid containing trusted and untrusted sensor node/relays. In the SCADA networks, various industrial and proprietary protocols will be integrated for cooperating together, such as TCP/IP standards and Common Industrial

Protocol (CIP) family. These protocols will be applied in the control process; however, these combined protocols are lack of protection mechanisms, which may introduce new and serious vulnerabilities for the security of the SCADA system [38]. The work in [39] also discusses the possible cyber vulnerabilities and their detection regarding the SCADA system.

2.3 Intrusion Detection System Background

Intrusion detection is defined as "the problem of identifying individuals who are using a computer system without authorization (i.e., 'crackers') and those who have legitimate access to the system but are abusing their privileges (i.e., the 'insider threat')" [40]. The intrusion detection system is "A computer system (possibly a combination of software and hardware) that attempts to perform intrusion detection" [41], Intrusion detection systems can be classified as host-based and network-based. Host-based systems make their decisions by the information obtained from a single host which is usually audit trails, and the network-based systems obtain data by monitoring the traffic in the network to which the hosts are connected [40]. Intrusion detection systems can also be classified into centralized intrusion detection system (CIDS) and distributed intrusion detection system (DIDS) by the way how their components are distributed [42]. A CIDS is the IDS that the analysis of the data will be performed in some fixed locations without considering the number of hosts being monitored [43], while a DIDS is composed of several IDSs over large networks whose data analysis is performed in a number of locations proportional to the number of the hosts. Components in the DIDS can communicate with each other or with a central server that facilitates advanced network monitoring. In the distributed environment, DIDS are implemented by applying co-operative intelligent agents distributed across the networks [44]. Most intrusion detection systems try to complete the attack detections in real time, but some

intrusion detection systems may not operate in real time, which may be raised by the natural reaction speed [45] or by the influence of the past record in the system [46].

The ideal characteristics of the intrusion detection systems are listed below [47].

- Work continually with minimal human supervision. However, since the IDS should be operated continuously by applying additional resources in the system, the detection and monitoring occur even when there are no intrusions, this may cause the resource usage problem and failure in addressing the normal operation.
- Fault tolerant: the IDS should be robust and be able to recover from system crashes either caused by accidentally or malicious activity; after a crash, the IDS should be able to re-heal its previous state and resume its operation in the shortest time.
- Subversion resistance: a significant difficulty should be set to prevent an attacker for disabling or modifying the system; the IDS should be able to monitor itself and find out whether its setting has been modified by an attacker. These three characteristics above may introduce the reliability problem to the IDS, since the IDS is implemented as separate programs, the programs of the IDS may be disabled or modified by the attackers, and the IDS will become useless.
- Impose a minimal overhead on the systems to avoid interfering with their normal operation.
- Be configurable to the security policies of the systems being monitored.
- Ease of development: this can be achieved by integrating different architectures and operating systems, through simple installation mechanisms, and by easy application and understanding from the operator.

- Adaptable to alterations in system and user behavior over time: the changes caused by new applications being installed, users changing from one activity to another, or new resources in system use patterns should be able to tolerate by the IDS.
- Ability of detecting attacks: considering the legitimate activity as an attack (false positives) should be avoided by IDS; flag any real attacks (false negatives) should also be prevented by the IDS, the IDS should report intrusions as soon as possible after they occur, and the IDS should be general enough for various types of attacks detection. However, since the information used by the IDS is obtained from the audit trails on a network, data needs longer path from the origin to the IDS with the danger of being destroyed by the attackers during the processing. Additionally, the IDS has the fidelity problem that the IDS may cause the misinterpretations when it is inferring the behavior and checking the data collected from the system [48], [49].

As one part of intrusion detection systems, the DIDS has specific advantages than the CIDS. For instance, the DIDS is highly scalable and can provide graceful degradation of service [47]. Therefore, during these years, more focuses are shifted to the construction of DIDS. Many distributed intrusion detection systems have been discussed in the literature, much of which is focused on applications in the area of the computer networks. For instance, in [50] a distributed community based intrusion detection system was proposed. This system was applied to detect security threats known as worms. The work in [51] proposes a dynamic, hierarchical, multi-agent-based DIDS that is designed to detect distributed attacks in communication networks. The work in [52] proposes a network-based DIDS that leverages an intrusion detection message exchange format to support digital signatures and increase the confidentiality of the entire communication system. An intrusion detection system which

uses autonomous modules was proposed to monitor the security of the networks in [53]. These models are successfully applied but limited to pure computer networks.

2.4 Computational Intelligence Algorithms in the Intrusion Detection Systems

As described in 2.3, the intrusion detection system is defined as a computer system for intrusion detection; it is therefore common for the IDS to imply the computational intelligence algorithms for the faster vulnerabilities detection. Some popular algorithms are applied in the IDSs as listed below:

- Artificial Neural Network (ANN)

ANN can be applied in supervised learning and unsupervised learning. One instance of supervised learning is Radial Basis Function (RBF)-based neural network is presented for the anomaly detection due to its simple architecture and quick learning ability in [54], it focuses on the parameters of the hidden layer and proposes a novel grid-based approach for data compression and clustering. In this algorithm, the Neural Network (NN) model is a traditional three-layer model and the two-stage learning process is employed in the training framework. The grid-based method divides the object space into numerical cells named hyper-rectangles or units unequally, which means the dimension of the dataset is divided into 10 unequal length intervals by applying the S-function which can amplify or decrease the results of the inputs.

The structure of the RBFNN is shown in Figure 2.1, in this structure, the inputs are attributes of the attack data; and it has a hidden layer and only one output node since this is anomaly detection.

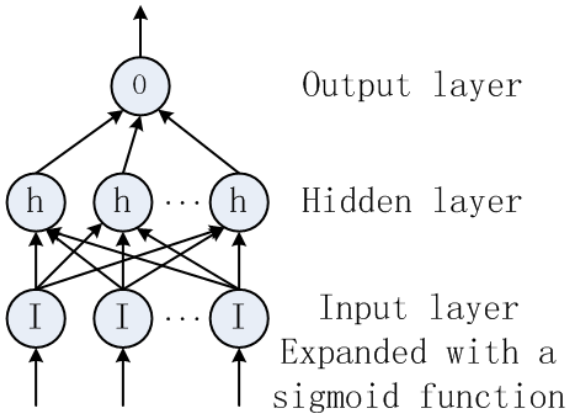

Figure 2.1: The Structure of RBFNN at the Detecting Stage

The instance of the unsupervised learning of the ANN is self-organizing maps (SOM) are single-layer feed forward networks, their outputs can be compressed into a low dimensional grid and they are the most widely used NN for anomaly detection tasks [55]. Lots of researches have involved SOM into the multi-layer agent detection system such as [56], which proposed a two layer detection framework. In the first layer a SOM was used for clustering the payload and compressing the data, the second layer which took the compressed data as input, would classify the data more accurately. Reference [57] uses the similar architecture which also consists of two layers. The first layer comprises of feature specific SOMs which perform like the detectors for separating the six TCP features, and the second layer consists of an integrating SOM which is used for combining the information of the six input features. Between the two layers, clustering algorithms such as fuzzy K-means clustering algorithm could be applied to increase the detection rate and decrease the false positive [58].

- Evolutionary Computation (EC)

EC is a field of research that concludes the ideas of evolutionary biology for optimization techniques. It includes genetic algorithms (GA), genetic programming (GP),

evolution strategies, and lots of other algorithms [59]. EC is also widely applied in the IDS, which constitutes and trains the software agents for a better detection performance.

In [60], GP is used for agents learning training, instead of saving a static intrusion types database in the system, the agents are required to learn the new intrusions and update their own knowledge-base.

GP combines a set of operators (functions) and a set of primitives that obtain the value of metrics in to the agent code, the code is then interpreted by an evaluator for information audition supplied by System Abstraction Layer (SAL). It is not difficult to discover from the figure that forward feedback occurs between SAL and evaluator. The process of the evaluation of the agent is presented by Figure 2.2 [60].

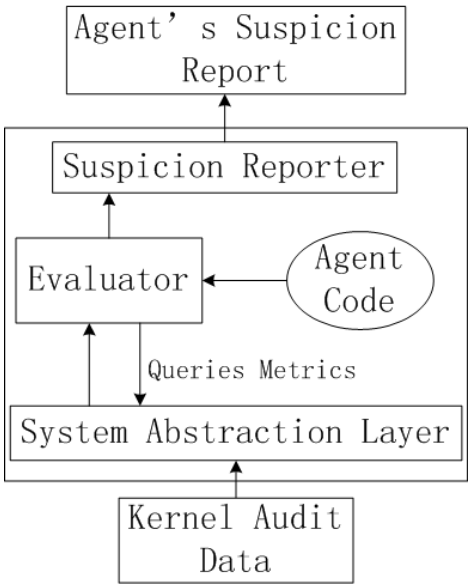

Figure 2.2: Agent Using GP [60]

The GP uses fitness value for evaluating the bias of the real intrusion situation and the predicted intrusions, by embedding the multiple Automatically Defined Functions (ADF) in 3 agents and monitoring the network timing of the source and destination, the suspicion reporter sends out a set of the suspicion value.

- Swarm Intelligence (SI)

SI is an artificial intelligence technique which involves the study of collective behavior in decentralized systems. It computationally emulates the behavior of the insects such as ants, bees, and termites to simplify the problems in the distributed environment. SI is popular in the application for optimization. Two popular SI methods in computational areas are Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) [59], [60].

In [61], a self-organized ant colony based intrusion detection system for intrusion detection in network architecture is presented. The positions are represented by a vector with the position r, and the orientation θ. The pheromone weighting function is expressed as in (1), which illustrates the transition rule between the points. The equation calculates the probabilities that the ant move to the point with the pheromone σ. The parameters δ and β represent the sensory capacity of each ant.

$$W(\sigma) = (1 + \frac{\sigma}{1 + \delta\sigma})^{\beta} \tag{1}$$

By combining different response thresholds for data updating, very high clustering rate has been achieved.

2.5 Intrusion Detection in Smart Grid

Since the smart grid network is a hybrid of the power system and a communication network, intrusions should be detected that concern either the physical power system or the communication network or both. The combination of these issues refers to the cyber-physical aspect of the smart grid. In [62], several secure smart grid models are proposed and approaches of critical infrastructure security for the smart grid are listed. In this thesis, cyber security of the cyber-physical power system is addressed, which considers the cyber components from both cyber and physical domains. For instance, phasor

measurement units and FACTS devices can be deemed as physical (hardware) components, but their cyber parts (e.g., embedded software) also can be intruded which will lead to various network traffic patterns. Our cyber security study also covers the cyber vulnerabilities from these physical devices which include both hardware and software components.

In the security system of the smart grid, it is crucial to choose an efficient communication topology and a reasonable communication standard as well as a robust attack classification algorithm. In light of this, many works about network topologies, standards, and classification algorithms were reviewed. In [10] and [63] the wireless mesh network topology was discussed and compared with other topologies. It was considered that the mesh network could supply a reliable, robust, and cost-effective topology for communication. In [64] the construction and parameters of wireless mesh networks in different environments were discussed. In [65] and [66] threats regarding the security of the wireless networks were presented and some solutions, such as secure routing protocols, were proposed.

The standards which are suitable for the wireless mesh topology are also studied. Zigbee is a low-cost and low-power network standard which can be widely deployed and utilized in the limited range of wireless mesh network environment [67]. The 802.11n Wi-Fi standard can enable a further range of communication with faster data transmission speeds than was previously cost prohibitive or impractical in wireless mesh networks [68]. WiMAX can also provide low costs and remote communication within the wireless mesh network [69].

One of the requirements of the SGDIDS method proposed is the ability to classify attacks efficiently and effectively through the use of a robust classification algorithm. In

order to accomplish this, a Support Vector Machine (SVM) is chosen as the classification algorithm because of its convenience of usage and high accuracy in classification [70]. Many experiments have also applied SVM algorithm on the KDD99 dataset [71] for data mining. In [72] an enhanced SVM model was created based on the important features of the training data. In [73] a radial basis function (RBF) was used as the kernel for the SVM algorithm in order to classify the KDD99 dataset. The work in [74] and [75] presents a methodology for applying SVM to the multiple classification of a large, scaled dataset. A popular SVM tool named LIBSVM [76] is applied for the classification of the intrusion data in this work as it provides good computational performance regarding attack detection.

In order to make a comparison with SVM algorithm, we also apply the artificial immune systems (AIS) as another classification algorithm. AIS are computational algorithms that emulate the mechanisms of human immune systems [55]. They involve learning, memory, and optimizing capabilities for conforming supervised and non-supervised computational algorithms. The primary advantages of AIS are that only positive examples are needed in the algorithms, and the patterns AIS has been trained with or learned can be explicitly examined. There are mainly four fields of algorithms which are derived from AIS: negative selection algorithms, immune network algorithms, danger theory, and clonal selection algorithms. In this application, the clonal selection algorithm is considered because of its flexibility. Its theory is used for emulating the basic process of an adaptive immune response to the antigenic stimulus. Only those cells that can recognize the antigens are allowed to clone and proliferate.

The clonal selection theory was first proposed by Burnet in [77] and numerical research has been completed by other scholars. For instance, De Castro proposed the Clonal Selection Algorithm (CSA) in [78], transformed it into the CLONALG algorithm in [79], the

Clonal Selection Classification Algorithm (CSCA) was reported in [80], and in [81] a dynamic clonal selection algorithm with the property of self-adaptation was proposed. In [82] Watkins proposed the Artificial Immune Recognition System (AIRS) algorithm which competes with CLONALG in terms of efficiency. The improvement of parallelization was described in [83].

Chapter 3

System Design

The method that we propose for integrating greater cyber-security into the smart grid communication scheme consists of three essential pieces: A three layer network infrastructure with multiple distributed modules, the use of SVM/AIS for attack classification, and an optimal communication scenario among distributed modules.

3.1 Network Architecture

When considering the large communication network that will exist in the smart grid, we propose a three-layer network composed of home area networks (HAN), neighborhood area networks (NAN), and wide area networks (WAN) that may be assumed as a reasonable infrastructure design.

The first layer of the network is the home area network (HAN) which consists of the service module (SM), the metering module (MM), and the HAN intrusion detection system (IDS) module. The SM will provide real-time energy cost and consumption data to the consumers while the MM (which includes the smart meters) will record the consumption of the energy in a consumer's home. The HAN IDS will, as expected, track and scrutinize both

incoming and outgoing communications in order to determine if any lapses in security are occurring.

The second layer is the neighborhood area network (NAN). The NAN is a large metering and controlling network which collects metering and service information from the multiple HANs that are geographically near each other. The NAN will consist of three components: The central access controller (CAC), the smart meter data collector (SMDC), and the NAN IDS. The CAC can be considered as the interface that manages the communication between HANs and the energy supplier or utility. The SMDC will be a wireless node that is in charge of the metering record of the whole community as composed by the neighboring HANs. All data that is flowing either in or out of the NAN will be passed through the NAN IDS in order to detect any possible security threats.

The final layer is the WAN. It provides broadband wired and wireless communication between the NAN, substations, other distributed grid devices, and the utility. This layer will consist of the energy distribution system (EDS), the SCADA controller, and the WAN IDS. The EDS will remain in charge of the energy and metering data distribution. The SCADA controller provides the utility or grid operator with distributed process control in order to manage the distribution grid elements. Since the large-scale control and metering data are crucial to the energy and service corporations (E&SC), a WAN IDS is required between the SCADA controller and the supplier in order to maintain security. A graphical depiction of all three levels of the proposed network can be seen in Figure 3.1.

Because of the recent implementation of standards-based wireless communication protocols such as IEEE 802.15.4 Zigbee, the widespread use of 802.11 protocols, the use of 802.16 WiMAX, and the need for distributed sensing technology, it can be assumed that large parts of each network layer will exist in the form of a wireless mesh network [10].

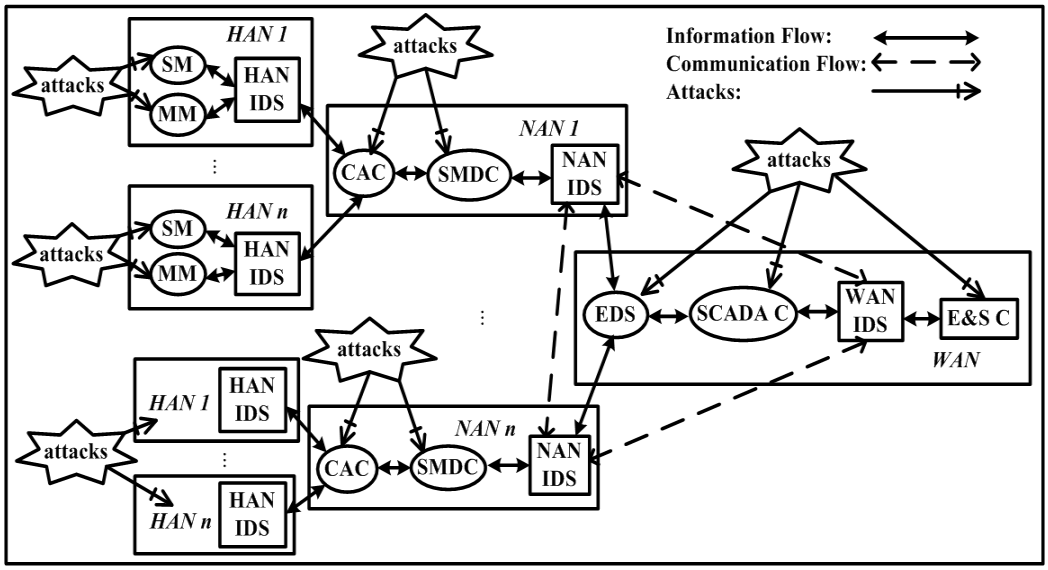

Figure 3.1: Three Layer Network Architecture.

The communication network topology chosen for the SGDIDS architecture is the wireless mesh network (WMN). The mesh network topology is adopted and is beneficial due to its redundant communication paths which compensate for natural failure as well as its scalable, dynamic and self-healing capacities [10]. Mesh networks also provide multiple, redundant communication paths that can compensate for the failures brought on by the break-down of communication between nodes due to failures, congestion, and other maladies. It is known that the majority of utilities have applied the wireless mesh system for smart grid components. The earth2tech's Katie Fehrenbacher estimated that 96% of the market is owned by wireless mesh networks when smart grid network installations are considered [42].

Since thousands of network standards are able to support the smart grid protocols and networks, requirements should be set for selecting the optimal network standards for the communication network of the smart grid, though these requirements will differ across the different layers. In the lower HAN and NAN layers the requirements include high network

capacity and data transmission rate. In the third layer, stability and reliability of the communication network is the main requirement.

In the HAN, the network standard chosen is 802.15.4 Zigbee. This protocol provides radio communication specifically suitable for personal and home area networks [30]. It has three available bands: 868-868.6 MHz with 1 channel for European utilities, 902-928 MHz with 10 channels for North American utilities, and 2.4 GHz with 16 non-overlapping channels for worldwide utilities. The 2.4 GHz band is chosen since it supports the data rate of 250 kbps as over-the-air data transmission which is higher than the data rates of other two bands. This band also has 16 channels. The maximum wireless capacity is calculated in (2) where C_{max} represents the maximum wireless capacity, r represents the data rate, and N_C means the non-overlapping channels. Thus, the maximum capacity of one data link under the 802.15.4 Zigbee standard is 4Mbps.

$$C_{max} = r \times N_C \tag{2}$$

In the NAN, 802.11n is chosen as the communication standard. Compared with other Wi-Fi standards, 802.11n increases the over-the-air data rates to a maximum of 300 Mbps on the 2.4 GHz band. This band is also shared with 802.15.4 Zigbee in the PHY layer [85]. The number of possible un-overlapping channels achieved is 23. The communication range is the largest when compared with the other 802.11 standards as the total range can be up to 250 m outdoors. By using (2), the maximum wireless capacity of links under the 802.11n standard is 6,900Mbps.

In the WAN, since the elements are mostly fixed and physically distant from each other, both wireless and wired networks will be applied to assist the availability and reliability between the communications of the smart grid elements. In our network, the large amount of data that is transmitted is sent via wired communication techniques such as power line

communication (PLC), while the control information is transmitted by a wireless network. The wireless standard chosen is the WiMAX 802.16-2009. This protocol is currently utilized with a 70Mbps data rate under the spectrally efficient orthogonal frequency division multiplexing (OFDM) mechanism. The most crucial reason for choosing WiMAX is that it can provide communication at a distance as large as 2 miles [69]. By applying the WMN topology and maintaining high data rate standards, the SGDIDS can provide a robust communication environment that is also congestion aware. For instance, when one access point of a smart meter is broken or flooded with traffic, the user can still read their metering information by accessing data from the other communication nodes or components. Also, by applying the broadband and high capacity wireless standards, SGDIDS is able to satisfy the requirement of transmitting a large amount of data.

Though the use of wireless communication technologies is a necessary aspect of the smart grid, these technologies introduce new vulnerabilities and security issues into the smart grid. Because of this, it is crucial to construct a cyber security strategy to protect the confidentiality, integrity, and availability of the data transmission in the smart grid. The inclusion of IDS technology provides a process of identifying the network activity that may lead to a compromise of security policy. It also monitors activities from access points while recording and preventing the suspicious activities which are marked as intrusions [72], [86].

3.2 Distributed IDS Modules

The HAN IDS is designed by combining multiple intelligent modules as shown in Figure 3.2. The Information Acquisition Module (IAM) collects the data packets of the energy consumption information and saves them in a matrix. The Data Segmentation Module (DSM) partitions the received data into proper-size segmentation files which are suitable for

the detection algorithm with the Preprocessing Modules (PM). The Analyzing Modules (AM) are used to detect suspicious intrusions. An AM is composed of three parts as shown in Figure 3.3. The intrusion data acquisition component saves the processed data from the PM and the intrusions are classified by using the trained SVM algorithm or the AIS algorithm. Finally the intrusion types, attack time, and the addresses are recorded and printed out by the output module. The whole detection process is, in the HAN, controlled by the Controlling module (CM).

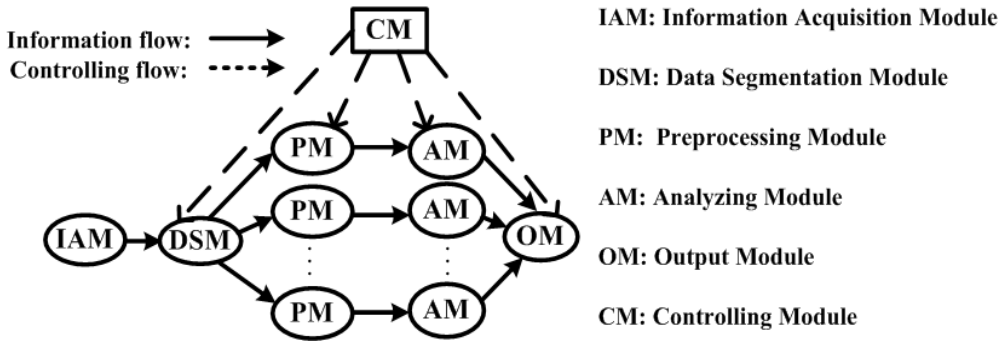

Figure 3.2: Home Area Network Intrusion Detection System.

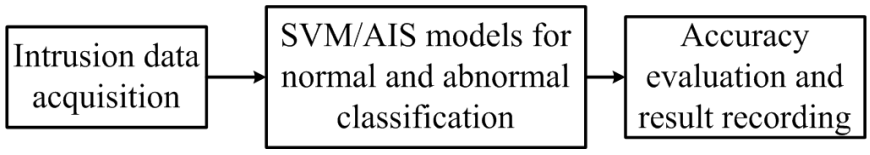

Figure 3.3: Analyzing Module Architecture

The IDSs that are designed for NAN IDS and WAN IDS are shown in Figs. 4 and 5. Each one is a combination of the best HAN IDS and other classifier models focusing on a better classification of the specific malicious attacks through the processing of massive amounts of data. For the attacks which are difficult to be classified in the current layer, the evaluation results will be checked in order to determine whether the data should be sent to

the higher layer for further evaluation. Also, in WAN IDS, a central controller is needed for controlling the NAN IDS in the sub-layer.

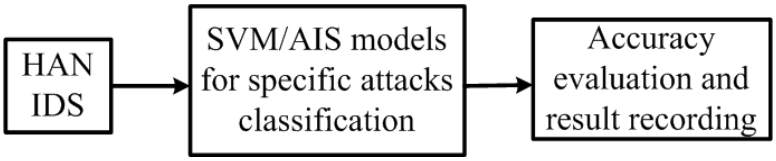

Figure 3.4: Neighborhood Area Network Intrusion Detection System

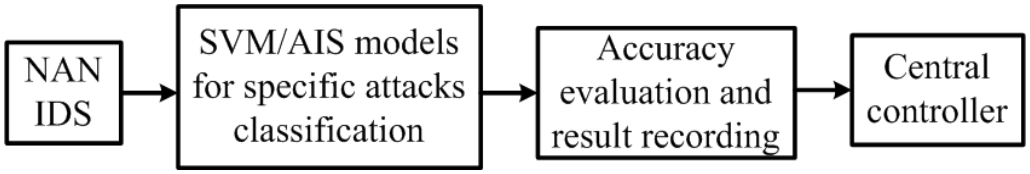

Figure 3.5: Wide Area Network Intrusion Detection System

3.3 Wireless Mesh Network Model and the Optimal Routing Algorithm

Here we use a directed graph, $G = (V, E)$, to illustrate the multi-hop multi-channel network [87]. Each node V can be considered as a smart grid component for measuring electricity consumption and managing communications with other nodes or higher layer components. Each edge, E, represents the wireless links between the nodes. In this graph interference amidst communications is not considered for simplicity. Only $C(e) \in C$ channels can be activated on any link e of the network where each channel is orthogonal to another. The data sent on each channel is transmitted by one and only one radio, i. Each channel of the data link has a capacity of $c_i(e)$. This is a constant showing the maximum data flow on that channel. The information flow currently transmitting on channel i of link e is denoted as $f_i(e)$.

In order to make sure the data flow can be transmitted in the channels smoothly, the constraints for the link or channel and all flows should be set appropriately. Suppose the scheduling variable $x_i^t(e)$ exists as $x_i^t(e)=1$ when data is transmitting on link channel i of link e during time slot t. Thus, the maximum number of the active channels is $C(e)$ during time slot t is:

$$\sum_{i\in C} x_i^t(e) \le C(e) \qquad e \in E \tag{3}$$

Also, we have λ sets of link and channel constraining pairs $(S_1, S_2, ..., S_\lambda)$ between each pair of nodes which are allowed to communicate to each other directly. In each pair, S_j means the constraining set belongs to channel i of link e. Each pair has a RHS constant, $\beta(S_j)$, according to the following constraint:

$$\frac{1}{\beta(S_j)} \sum_{(e,i)\in S_j} \frac{f_i(e)}{c_i(e)} \le 1, \qquad j \in \{1,2,...,\lambda\} \tag{4}$$

The nodes communication's with each other consist of M source destination pairs (shorted as commodity pairs) denoted as $(s(m), d(m))$. In each pair the flow allowed on channel i is set to be $p_i^m(e)$ and the total flow, $r(m)$, transmitted between the pair can be calculated as $\sum_{e=s(m)} \sum_{i\in C} p_i^m(e) = r(m)$ according to the constraint

$$\frac{1}{\beta(S_j)} \sum_{S_j} \frac{\sum_{m\in M} p_i^m(e)}{c_i(e)} \le 1 \tag{5}$$

The primal-dual algorithm is applied to find the optimal path for communication flows. Set SF as the maximum scaling factor where its value denotes the total slack capacity needed in the network. Because of the constraints above, it is required that SF should not be smaller than 1. The primal dual algorithm is described as Figure 3.6:

$$
\begin{aligned}
&w(j) = \delta, j \in \{1,2,..., \lambda\} \\
&b = 0 \\
&While \sum_{j=1} w(j) < 1 \\
&\quad For\ m = 1,2,...,M \\
&\quad\quad r = r(m) \\
&\quad\quad While\ r > 0 \\
&\quad\quad\quad Set\ \Delta w(j) = \frac{\sum_{j \in S_i} \frac{\alpha(j)}{\beta(S_j)}}{c_i(e)} \\
&\quad\quad\quad P_{\min}(s(m),d(m)) = \min \sum_{m \in M} \Delta w(j) \\
&\quad\quad\quad \mu = \min(f(P_{\min})) \\
&\quad\quad\quad \sigma = \min\{r,u\} \\
&\quad\quad\quad r = r - \sigma \\
&\quad\quad\quad f_i(e) = f_i(e) + \sigma \\
&\quad\quad\quad w(j) = w(j) * (1 + \frac{\varepsilon\sigma}{f(P_{\min})}) \\
&\quad\quad endWhile \\
&\quad endFor \\
&\quad b = b + 1 \\
&endWhile \\
&g = \max\left(\sum_{j \in S} \frac{f_i(e)}{c_i(e)}\right) \\
&SF = \frac{b}{g}
\end{aligned}
$$

Figure 3.6: Primal Dual Algorithm [88].

In this algorithm $w(j)$ is the weight of the constraining pair of link i channel e and $\alpha(j)$ is the weight of each set S_j whose value is calculated by the distance between two nodes. When searching for the optimal route, all the routes starting from the source node and ending with the destination node will be counted and the accumulation of the constraining pairs' weights will be calculated. Then the route which has the lowest value of weight accumulations will be labeled as the shortest path. The desired data flow, $r(m)$, is distributed to the corresponding node, and this source destination pair, which is also called a commodity pair, can find the optimal route with the shortest distances and link weights. By using this algorithm, the problem of finding the optimal path and the flow distribution is settled. Through the application of the primal-dual algorithm to find the shortest path of communication, a communication network could be provided to satisfy the requirement of efficient power consumption in the smart grid.

3.4 SVM for Attack Classification

Machine learning via SVM has previously been found to exhibit improved performance and is a powerful tool for the classification of data [89]. SVM is a type of machine learning technique that attempts to successfully classify sets of data by leveraging two basic principles: Large-margin separation and kernel functions. Large-margin separation refers to the idea that when classifying data it is sensible to draw a line of separation in such a manner that the distance between that line and the closest data point on either side of that line is maximized. Kernel functions are algorithms or functions that calculate the similarity between two points and must be used if non-linear classification boundaries are required. Generally speaking, as the dimension of classification increases hyper planes are used to separate the

data instead of lines. This is completed by mapping the data to a different space (also using a kernel function) where a hyper plane is able to classify the data.

In order to achieve a classification based on these ideas, a convex, quadratic program must be solved that is, in the case of data that is non-separable, of the form:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \qquad (6)$$

$$\text{such that} \quad y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \qquad (7)$$

$$\xi_i \geq 0 \qquad (8)$$

where w is the weight vector, C is a value that and controls the variance between margin maximization and error minimization, ξ_i is a set of slack variables that measure constraint violation, y_i denotes that there is a unique constraint for each necessary classification i, b is the bias, x_i is a training vector, and $\phi(x_i)$ is the kernel function that maps the training or testing vector to a different space [88], [90]. The formulation is the equivalent of maximizing the margin while also minimizing error.

Often it is also necessary to classify data into three or more categories. Two methods are typically used to perform this classification: the one-against-many and the one-against-one approach. The former approach is simpler and examines one category at a time while merging the remaining categories into a single category. This allows for simple, binary classification. The latter approach is more computationally intensive though more accurate

as it creates $k(k-1)/2$ models in order to classify all data where k is the number of classifications required.

3.5 Clonal Selection Classification Algorithm for Attack Detection

The clonal selection principle describes the process of the reaction between the immune system and antigens. During the reaction, lymphocyte cells are created so that the unknown antigen is better recognized. The cells start their proliferation through cell cloning and differentiation that is based on mutation. The activation of the B-cell evaluates its binding to the antigen. The number of clones produced by the activated B-cell depends directly on its level of activation. The higher level the activation has, the more clones will be produced. This fact leads to faster immune adaptation since B-cells with low affinity are differentiated through high mutation rates. This may also increase the affinities of the following generations. The B-cells with poor affinities after the differentiation process are effectively eliminated. The elimination of less qualified B-cells and the affinity-proportionate cloning of B-cells emulate the natural selection process, meaning that only the cells with highest adaptation to the environment are selected and regenerated [91].

In the area of AIS there has been evidence that the nature of the immune system can be realized and applied in network intrusion detections [92]. In this chapter we apply two clonal selection algorithms, CLONALG and AIRS2Parallel, for the generation of classifier models.

3.5.1 CLONALG Algorithm

The CLONALG algorithm is successfully applied in many complex problems with a low complexity structure. When being applied in classification problems it provides promising

accuracy. CLONALG was proposed by De Castro, named as the clonal selection algorithm (CSA), and later renamed as CLONALG. This algorithm takes a population of antibodies and then creates a population which is more sensitive to the antigens by exposing the antibodies to the antigens for several generations. As described in [78] and [80], the basic CLONALG algorithm achieves this by repeating the maintenance, selection, cloning, and mutation of the antibodies until they reach a threshold of sensitivity to the antigens.

Compared with the AIRS algorithm, CLONALG has a lower complexity and a smaller number of parameters that influence the classification accuracy.

The main algorithm for a single generation of CLONALG is described as follows:

1. Randomly generate an initial population of antibodies (A_p), which is composed of two subsets: A pool of the memory cells (M_p) and the reservoir pool (R_p). Also create a set of antigenic patterns, A; and

2. Select one antigen A_i from A, measure the affinity between A_i and every member of A_p, and pick n antibodies which have the highest affinities;

Affinity is the value of the similarity between two sets of data. A simple approach to measuring affinity is the Hamming distance between two sets of strings with equal length. The Hamming distance is calculated by using (9):

$$D = \sum_{i=1}^{m} \eta \qquad \text{where} \eta = \begin{cases} 1 & A_{Pi} = A_i \\ 0 & A_{Pi} \neq A_i \end{cases} \tag{9}$$

where D is the Hamming distance, A_{pi} is the antibody, A_i is the antigen, and m is the length of the data string.

3. Regenerate the clones of the n antibodies with the highest affinities and place them in a new population C. The number of clones is decided by the values of the affinities;

4. Mutate the clone population C at a rate inversely proportional to their affinity to produce a mature population C`;

5. Re-evaluate the affinity to each member of the population C` and select the antigen with the highest affinity as the candidate memory cell of the memory cell set M_p. The member of the memory cells is updated when the affinity of C_i is greater than the current M_{pi}; and

6. Replace the antibodies with low affinities in R_p with those of high affinity. Repeat two to six until all the antigens are involved and present it to the antibodies. This completes one generation of CLONALG is completed. Figure 3.7 provides a complementing diagrammatic representation of the workflow of the CLONALG algorithm.

After G generations of immune training, a matured memory pool is created as the result of the CLONALG algorithm. The elements in the memory pool can be applied to different problems such as pattern recognition, optimization, and the Travelling Salesman Problem (TSP).

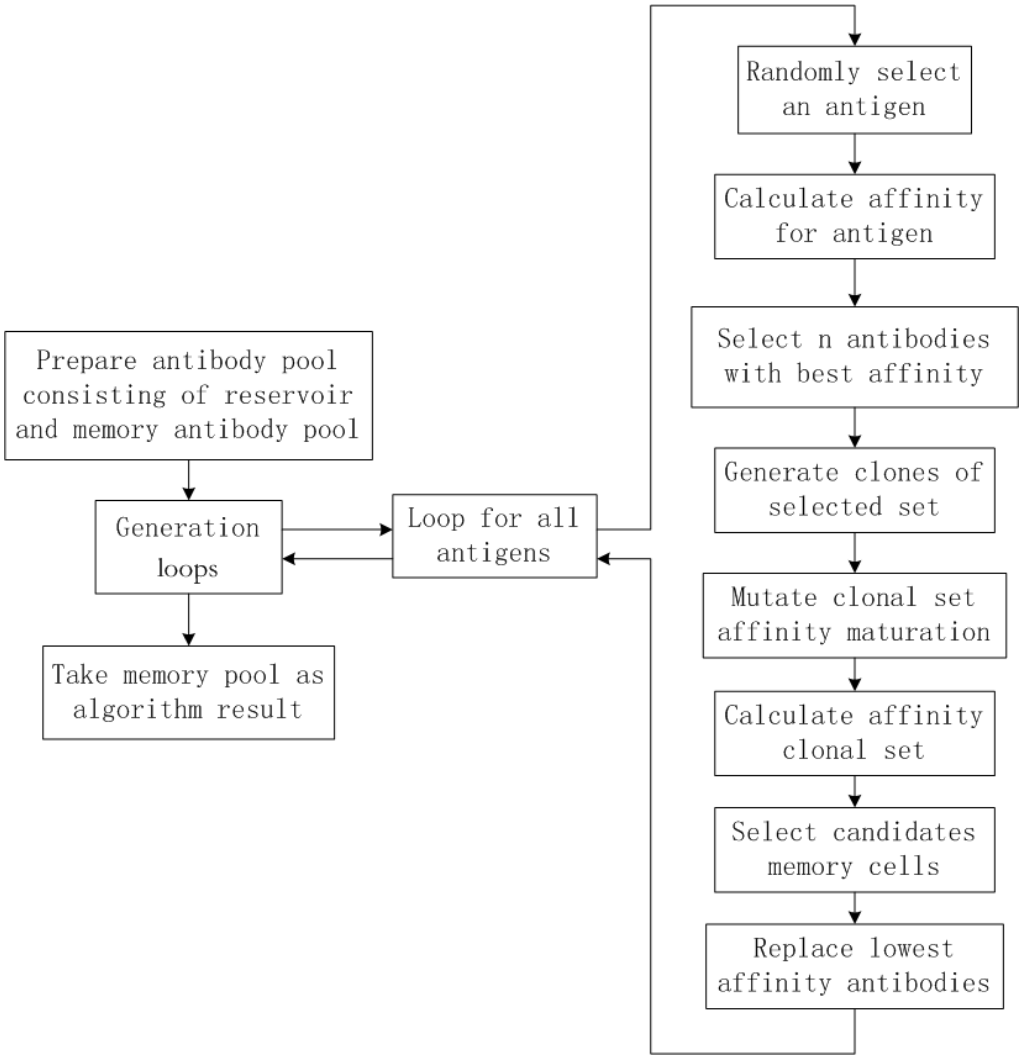

Figure 3.7: Diagram of the CLONALG algorithm [80]

3.5.2 AIRS2Parallel Algorithm

The artificial immune recognition system (AIRS) algorithm is one of the immune-inspired supervised learning algorithms which include clonal selection, affinity maturation, and affinity recognition balls (ARBs) [83]. In the experiments of [93] it has been concluded that AIRS has the highest value of classification accuracy. The AIRS algorithm is a cluster-based approach especially for data classification that optimizes the location of cluster centers. This is different from CLONALG which is an unsupervised learning

algorithm. But similar to CLONALG, AIRS also focuses on the development of the memory cells over multiple generations. There are five steps in the AIRS algorithm: Initialization, antigen training and competition, memory cell selection, and classification of the dataset. The second to fourth steps are repeated iteratively.

First, the dataset is normalized and the affinity threshold variable is calculated. Then two different populations, the artificial recognition ball (ARBs) pool and the memory cell pool are created. In the antigen training step, each antigen is required to be exposed to the memory pool. The best memory cell for the memory pool is selected when the stimulation value is maximized during the stimulation process. This cell is then applied for the affinity maturation process where it is cloned and mutated based on its value of the affinity. The clones are then added to the ARB pool. The competition then starts and ARBs which have low resources are discarded from the pool until the stopping criterion is satisfied. The ARB which has the highest stimulation value is selected as the candidate memory cell and is added into the memory cell pool. These steps are repeated until all the antigens are tested and then the classification will take place by performing the k nearest neighbor search in the memory pool [94]. Figure 3.8 illustrated the life circle of the AIRS algorithm.

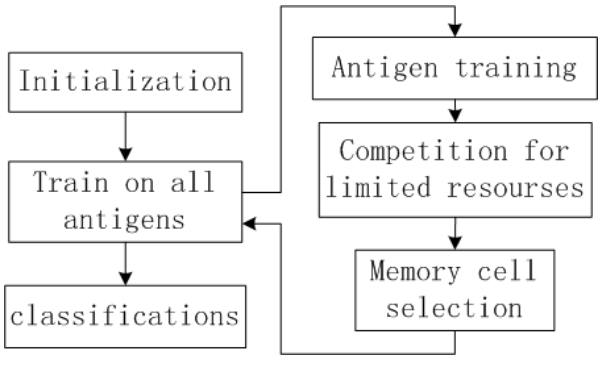

Figure 3.8: The life cyber of the AIRS algorithm [95]

The AIRS2Parallel algorithm may also increase computational efficiency [96]. The difference between this algorithm and the AIRS algorithm is that the dataset used in AIRS2Parallel are departed for the parallel calculation and the combination of the memory cells into a single pool for classification. By using this approach, both computation speed and the accuracy of the classification are increased.

3.6 Dataset and Preprocessing

In the proposed models, the SVM and AIS algorithms will be used as the two methods of attack classification. Each IDS at the HAN, NAN, and WAN level will incorporate these methodologies in order to detect attacks and stave off intrusions. In order to accomplish this, we have incorporated SVM using LIBSVM and CLONALG/Airs2Parallel using the Weka plug-in [97]. These tools are use with an improved and simplified version of the KDD Cup 1999 Data set (KDD99) [71] called the NSL-KDD dataset [98].

The NSL-KDD dataset is a short form of KDD99 which was gathered at MIT Lincoln Labs. KDD99 is the most widely utilized dataset for the evaluation of the anomaly detection as it is an extremely large data set that has a huge number of redundant records of network traffic. However, because of the redundant and repeating records of DoS attacks and the biased distribution of the four types of the attacks, the accurate classification of the U2R and R2L becomes difficult. To address this problem, we use NSL-KDD as the dataset for training and testing. The NSL-KDD dataset discarded the redundant and duplicated records so that the training time is significantly reduced. In total, the training data consists of 125,937 samples consisting of 41 features and 22 separate types of attacks and the testing data composes of 22,544 samples of records. The attacks can be divided into four separate categories including Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L),

and Probing/Scanning. This dataset can be considered as an abbreviated form of possible attack scenarios that are applicable when considering the smart grid's communication network.

Smart grid can be considered as a combination of the computer network and the power system with some new characteristics. Thus, some communication scenarios in the current computer network can be applicable to the smart grid. For instance, some attacks occurring in the computer network can find their counterparts in the smart grid such as the block of the communication traffic and the theft of the data. Although the KDD 99 dataset used in the SGDIDS is based on the communication scenarios in computer networks, the features of the record samples such as the service duration time can also be applied to the smart grid communication network scenario. Besides the network traffic patterns, other features of the power system itself (i.e., the current-carrying part) such as the voltage or current values, power flow values, and phasor measurements can be incorporated to enhance the dataset. Based on the variations of power system characteristics under different cyber attacks, the dataset used for classifier training and testing can be developed.

Concerning SVM, our model uses a Gaussian Radial Basis function for the kernel [85]. The models are trained using a partition of the NSL-KDD dataset. Because of the hierarchical three-layer structure of the SGDIDS, the distribution of training data differs layer by layer. As IDSs in the HAN layer have a limited chance of suffering from most types of attacks, we use training data that has a higher concentration of Normal and DoS, Probing attacks. The NAN IDSs data differs as its focus is on less common attacks like the R2L attacks. The WAN IDS will focus more heavily on U2R attacks. Because of these differences, training data that have different concentrations of attack types are used at each level.

The training and testing data used is prepared for classification by using the pre-scaling approach to create a format suitable for use with the SVM methodology and the LIBSVM library [13], [99], [100]. In large scale data mining it is very important to make the dataset sparse. In order to accomplish this, all non-numeric data (protocol types, service types, and flag types) are mapped to numeric numerals. As an example, in this study the feature called 'attack type' is changed from the values DoS, R2L, U2R, and Probing to the values 1, 2, 3, and 4 respectively. After this step, all numeric data types are scaled to a range between 0 and 1. This is done by dividing each specific value (V) by the difference of the maximum (F_{max}) and minimum (F_{min}) value for that feature according to (10). All of these values are then stored in a file in the format specific to LIBSVM.

$$\frac{V}{\left|F_{\max} - F_{\min}\right|} \qquad (10)$$

In CLONALG and AIRS2Parallel, training and testing data sets used are preprocessed for classification by using the NormalizeMidpointZero filter in Weka for pre-scaling. This increases the accuracy of the classification. After this step, all numeric data types are scaled to a range between -1 and 1. All of these values are then stored in a file and trained by using the Weka GUI platform.

All simulations are run using Matlab on a computer with Ubuntu 10.04, two Intel Xeon 5504 quad core processors, and 16 GB of RAM.

Chapter 4

Simulations Results and Analysis

4.1 Distribution of Possible Attacks in Different Layers of SGDIDS

In the smart grid network, the threats launched are different from those found in a typical computer network. In [100], forms and approaches of the attacks in the smart grid scenario are surveyed and discussed. It is found that the specialized layout, design, and communications of the smart grid network expose it to more possible vulnerabilities than a typical computer network. As an example, the AMI may suffer from cyber vulnerabilities related to routing, configuration, name resolution, or authentication protocols while also suffering from physical (hardware) vulnerabilities and encrypted key access.

When compared to the attacks in computer networks, attacks in the smart grid environment should also consider the motivations of its attackers. These motivations can be categorized into five areas: Curiosity for information, motivated attacks, unethical power theft, power consuming information theft, and financial motivations for economic benefits [57].

While these motivations are important, our model considers all attacks and motivations to be encapsulated in the NSL-KDD dataset. These attacks are also illustrated in Fig. 3.1. Based on literature regarding the cyber security of the smart grid, it can be assumed that diverse attacks occur in different layers of the SGDIDS.

In HANs, wireless probing of consumers metering data and the stopping of data transmission between meters may occur. These types of attacks belong to the categories of Probing and DoS attacks respectively. In NAN, Probing attacks may occur in order to steal the control information of the CAC or to disable the connection between the metering modules and the CAC can be constructed by the DoS attacks, what's more, the data security parameters or controlling information could be reset or altered by the U2R attacks. In WAN, besides the DoS attack as disruption of the metering data transmission to the EDS or E&S C, and the Probing attack as large scale disclosure of customer's metering data, the components in WAN may also suffer from the U2R and R2L attacks such as the alteration of metering and controlling data in the EDS and E&S C, or broadcasting of control commands from the SCADA [8], [57]. In a word, the types of attacks increase in the higher layers of the architecture.

The IDS that is placed into the communication network will detect the existence of and classify the types of these intrusions.

4.2 Grid Topology and Communication Routing Optimization

The wireless mesh communication scenario that uses SGDIDS is simulated using MATLAB as it is a powerful tool for simulating a complex communication system. Figure 4.1 shows the simulation of the hierarchical, three layer smart grid communication network proposed in Figure 3.1. As can be seen, the network is composed of components by nodes

with nodes 1 to 20 in the HAN, nodes 21 to 23 in NAN, and node 24 in the WAN. In order to ensure the security of each component, every node has an IDS belonging to its corresponding network, It means the HAN IDSs are embedded into nodes 1 to 20 in the HAN, and the NAN IDS and WAN IDS are combined in the nodes 21 to 23 and 24 respectively in NAN and WAN.

In Figure 4.1, a mesh network with 20 HAN nodes which are scattered randomly in a 500×500 square is generated. These nodes representing the SM or MM compose several HANs. According to the rules of the communications among the nodes in a wireless mesh network, a node can directly communicate with another node by the link between them as long as it is located less than a specified distance away. Here the direct communication distance as set to about 100m since the longest communication distance of Zigbee is 150m. Remote node pairs that need to communicate with each other will use the multiple-hop approach by sending their communications through other nodes. The medium sized nodes (21 to 23) represent the CAC or SMDC components in each NAN. These components can directly communicate with the nearest nodes in HAN by using the 802.11n standard. Remote nodes in HAN that need to communicate with other nodes in the NAN can send the information through other HAN nodes. Since one NAN node may route the metering and detection information from multiple HAN nodes, the capacity of the each node and link in the NAN is set higher.

Adding the largest node (24) into the WAN completes the entire 1000×500 network. The WAN node can be considered as an EDS, a SCADA Controller, or the E&SC. As the communication between the WAN node and NAN node is completed across a larger distance, the majority of power for communication is consumed at the intermediate nodes. As such, it is necessary to make sure that there is enough power available at these

intermediate nodes. The wireless communication standard between a NAN node and the WAN node uses the WiMAX protocol. The link between a WAN and NAN has only one wireless channel in this study. The wired communication standards are applied for metering data transmission because of the larger distance and the great amount of data.

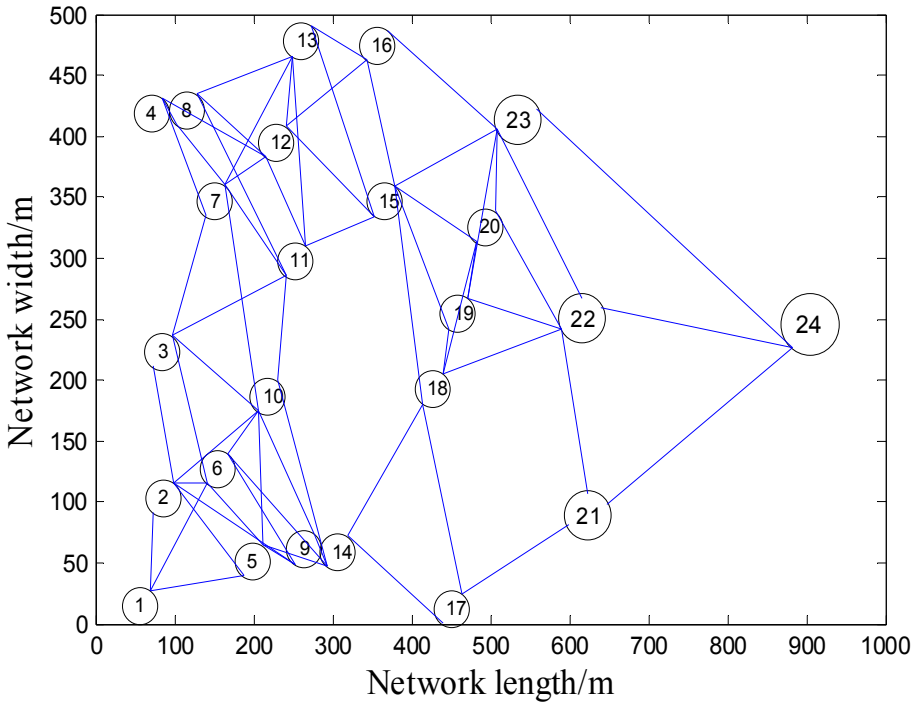

Figure 4.1: Smart Grid Distributed Intrusion Detection System Topology

The source and destination node pairs are randomly picked from HAN, NAN and WAN nodes. The source nodes are tasked to send out packets to the destination nodes directly or by using multiple hops. In this simulation, five commodity pairs are set with each pair labeled using one color. The corresponding route is also labeled with the same color that is used by the commodity pair. If the node is both used as the source and the destination, its color would be overlapped. The shortest routing for the commodity pair is found by using the primal dual algorithm described in Section 3.3. In order to more realistically simulate the data transmission, the number of channels and the data rate is randomly set on each link under the rule of the network constraints. This means that in the HAN a maximum 16

channels can be opened simultaneously on one link. In the NAN the highest number of channels is 23. Figure 4.2 illustrates the communication routing between the commodity pairs and the optimal routings are recorded in Table 4.1.

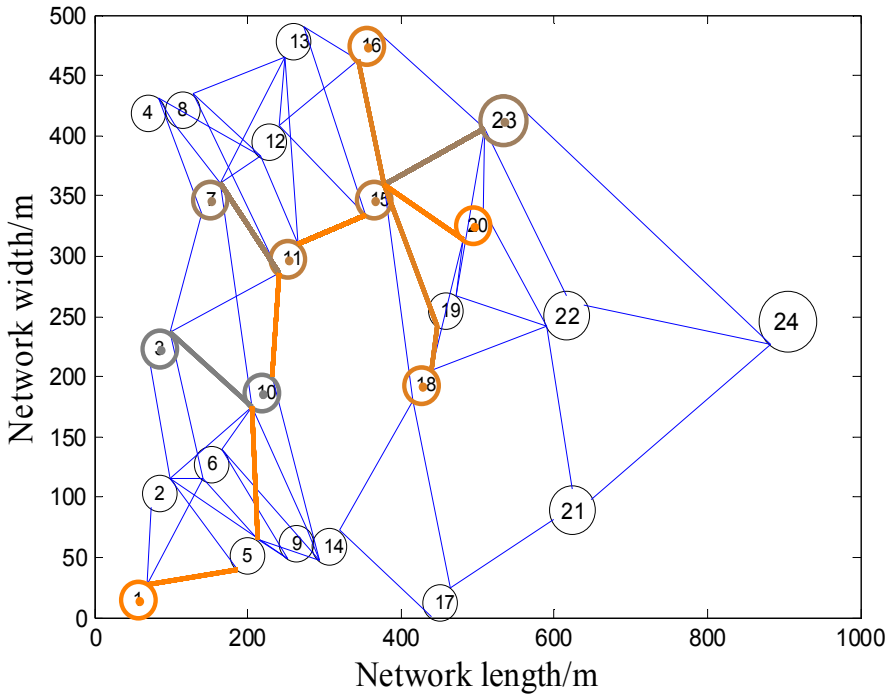

Figure 4.2: Optimal Communication Routing Selection in the SGDIDS

In Table 4.1, it can be seen that the route chosen by commodity pair (16, 18) uses nodes 15 and 19 as hopping nodes. Although the distance between node 15 and 18 is shorter than the distance from 15 to 19 and 19 to 18, the link between 15 and 18 has a higher weight value since fewer channels are open on this link. This is an example of how using only the shortest path may bring about communication congestion. Thus, by applying the Primal Dual algorithm, optimal routing of the communication could be achieved.

Table 4.1: The Routing Table of the Communication among Nodes.

Source node	Intermediate nodes				Destination node
3					10
7			11	15	23
11					15
16			15	19	18
20	15	11	10	5	1

4.3 Intrusion Detection with the SGDIDS

The detection result is presented in Figure 4.3. When the packets arrive at the destination node, the destination node will use its IDS to extract the 41 features of the KDD99 from the communication. Some features that are examined include the duration of the node's connection and the protocol type of the connection. The DIDS will then use its classification mechanism (SVM or AIS) to detect whether this set of features represents normal communications or an attack.

The classifiers are trained by using a specific portion of the NSL-KDD dataset in LIBSVM and the Weka plug-in for AIS algorithms. As these algorithms are nonlinear algorithms, the accuracy of the testing doesn't depend on the size of the training dataset. Rather, it depends on the proportions of the normal and the malicious records.

Our SVM model uses a Gaussian Radial Basis function for the kernel [74], a value of 2^{15} for *C*, a value of 2^{3} for ξ, and a value of 0.1 for ε as the parameters. The choice of the cache size (a parameter particular to LIBSVM) is 1,600 MB because of the large scale of the training dataset.

Regarding the two AIS, the classification performance of the AIRS model is influenced by eight parameters and the performance of the CLONALG is influenced by six parameters [101]. The optimized values of the parameters which lead the algorithm to the best

classification results are listed in Table 4.2. The choice of the cache size for operation of Weka is 2,500 MB in order to prevent the out of memory errors.

Table 4.2: The Influencing Parameters and the Optimized Values

CLONALG		**AIRS2Parallel**	
Parameters	**Values**	**Parameters**	**Values**
CF (Clonal Factor)	0.9	ATS (Affinity Threshold Scalar)	0.4
APS (Antibody Pool Size)	50	CR (Clonal Rate)	10
SPS (Selection Pool Size)	20	HR (Hyper-mutation Rate)	2
TR (Total Replacements)	2	kNN (Number of Nearest Neighbors)	3
NG (Number of Generations)	10	IMPS (Initial Memory Cell Pool Size)	50
RPR (Remainder Pool Ratio)	0.1	NIAT (Number of Instances for Affinity Threshold)	-1
		ST (Stimulation Threshold)	0.5
		TR (Total Resources)	150

We distribute twenty HAN IDS modules, three NAN IDS modules, and one WAN IDS module into the network nodes using both SVM and AIS algorithms. From the possible attacks analyzed in Section 4.1, the HAN IDS modules includes models for Normal communications, DoS attacks, and Probing attacks since these two types of attacks are more likely to occur. The NAN IDS modules include the best testing model of these HAN IDSs and the model trained for classifying the U2R attacks. For the module in the node of WAN IDS, besides the best HAN IDS model and the best NAN IDS model, it has the third model which is used for classifying the R2L attacks.

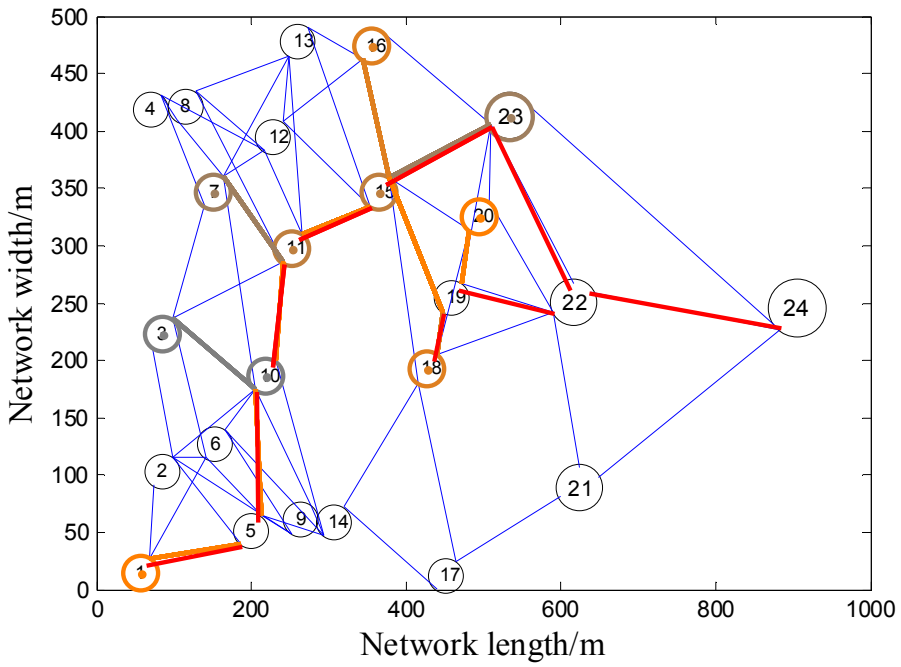

Figure 4.3: Intrusion Detection Routing in Home, Neighborhood, and Wide Area Network.

As the process of simulating packet transmission and gathering data (such as packet type, connection duration, etc.) is complex, we have simplified it by including features of KDD99 dataset as a payload. It is this payload data that is classified as either normal or malicious.

If a source node has a communication with a destination node and the IDS in the destination node cannot clearly detect what kind of communication those features belong to, the features are then packaged and sent to a higher layer IDS for classification. Because of the limited possibility of types of attack samples, the IDSs in the lower layer can only detect limited types of intrusions as was stated in section 4.1.

In any case, if the HAN IDS is unable to classify network traffic, then the data will be packaged and transmitted to a NAN IDS. If the NAN IDS is unable to classify traffic, it will first attempt to contact the closest NAN IDS for classification. The HAN IDS does not have this mechanism because of its smaller communication capacity and the limited

detection ability. If the NAN IDS fails at classifying the data, it will then be sent on to the WAN IDS as it has the highest classification accuracy and ability. In Figure 4.3 the search pattern for traffic classification is detailed. For example, the red line between nodes 18 and 19 represents the detection IDS as it is switched from HAN IDS to NAN IDS. The detection procedure is illustrated in Figure 4.4.

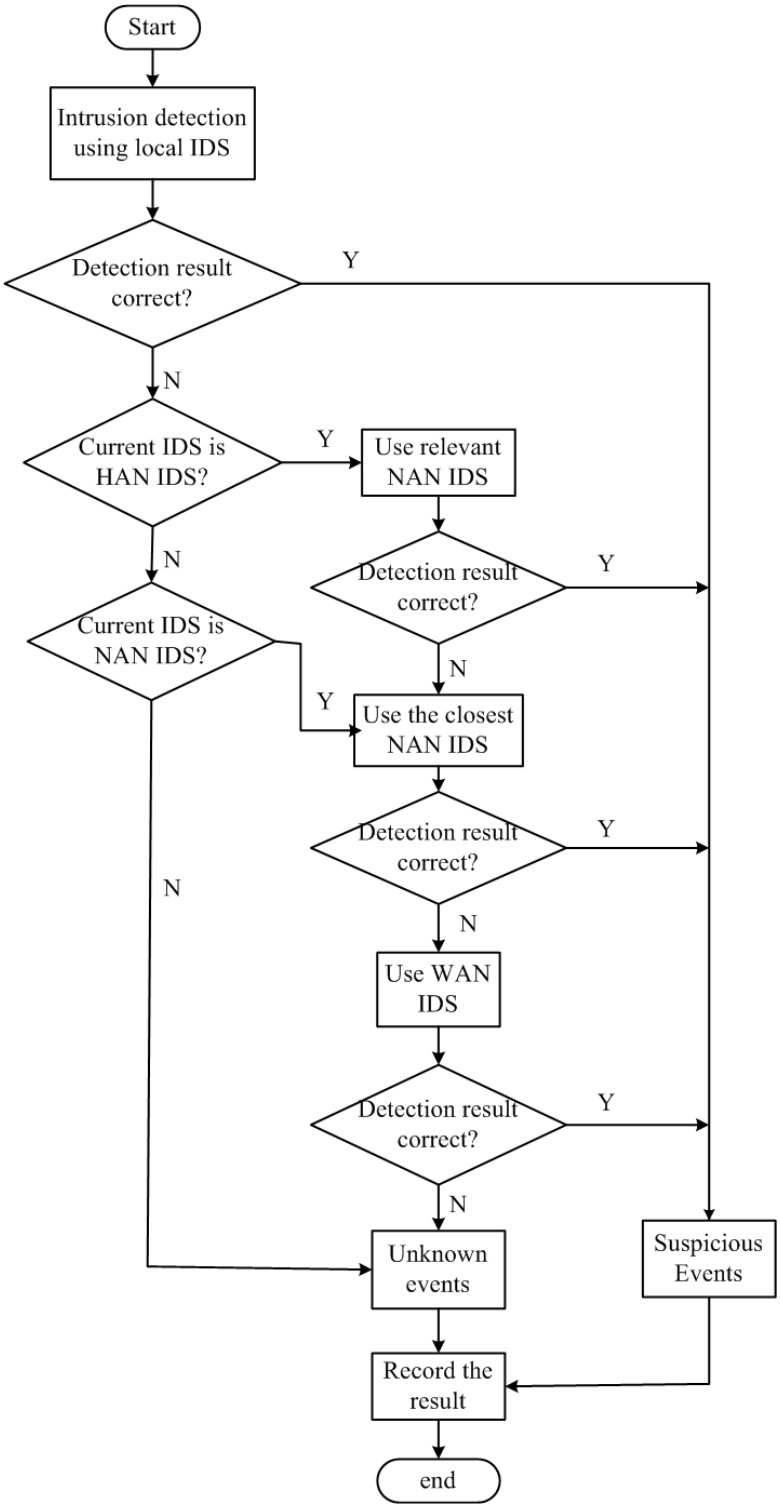

Figure 4.4: Detection Procedure in the SGDIDS

The intrusion detection routing table is illustrated in Table 4.3. The first column shows the destination nodes in part 4.3 that use the local IDS to make the detection and the last

column represents the final IDSs used for the classification. The nodes between the destination nodes and the final nodes failed to perform a successful classification. The rule of optimal routing is also followed here. The first row of Table III illustrates the cooperation that can occur between the IDSs at multiple levels. In this case, the HAN IDS in node 10 then attempts to classify the communication as benign or malicious but is unable to do so. Because of this, node 10 packages and passes the communication up the network hierarchy to node 23 through the optimal routing of (10, 11, 15, 23). The NAN IDS in node 23 then attempts to classify the communication using its own AM, but it also fails to do so. The communication is once again packaged and passed to node 22 which has another NAN IDS. As this IDS cannot successfully classify the data, it will pass it up the hierarchy to node 24 with a WAN. If a successful classification is made, the classification will be transmitted down the hierarchy to the source node. The second and third rows show that the classifications succeed locally.

Table 4.3: The Detection Routing Table of the Communication among Nodes.

Destination node	**Intermediate nodes**				**Final node**
10	11	15	23	22	24
23					23
15					15
18			19		22
1	5	10	11	15	23

4.4 Results and Discussion

In the real smart grid communication network environment the transmission of packets will be continuous. This is simplified in our simulation as it is run a total of 50 times. Every

time a record sample is randomly abstracted from the test dataset and sent to the IDS in the destination node, and the classification occurs as described in Section 4.3, then a notice is displayed on the screen showing whether there is any threat to the network. The amount of time spent in classification, a count of the normal communications, a count of the total attacks, and the number of classification failures will be recorded and illustrated on the screen as well. If a classification cannot be properly made, a "detection failed" note will be posted and the total accuracy will be decreased.

In the simulation there are 113 normal communications, 82 DoS, 34 R2L, 2 U2R, and 19 Probing attacks launched to the SGDIDS. The accumulation of the classification results and the classification failures using the SVM, CLONALG and AIRS2Parallel algorithms from the simulation are recorded in Figure 4.5. The first column shows the number of the normal communications followed by the accumulations of DoS, R2L, U2R, and Probing attacks. The final column shows the number of times that classification fails. From Figure 4.5, it can be found that the performance of the SVM is better than CLONALG and AIRS2Parallel algorithm at finding the R2L and U2R attacks. The classification failures of AIS algorithms are mostly due to insufficient attack samples in the training dataset. In particular, in roughly 126,000 records of training dataset only 995 are R2L and 52 are U2R. However, the SVM can classify them by using the appropriate proportion of data in the training dataset.

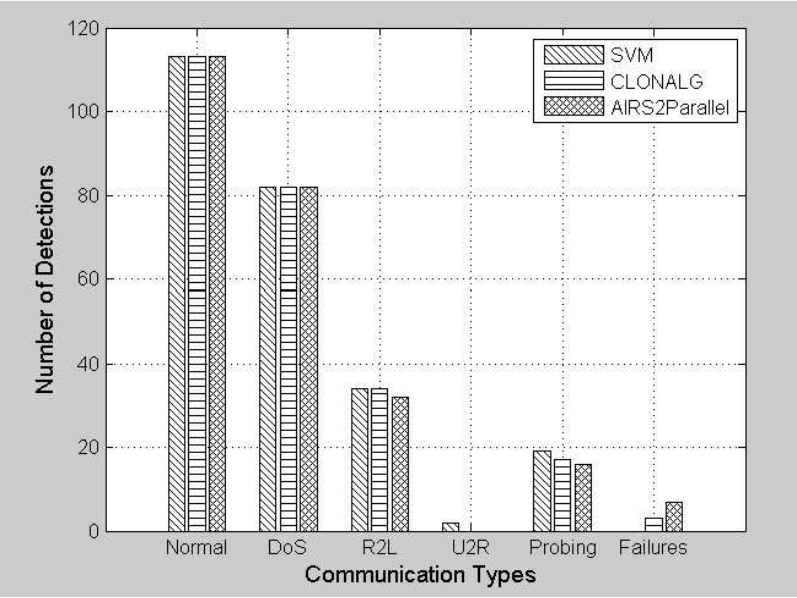

Figure 4.5: Results Using SVM and AIS Algorithms

In order to achieve a better comparison of these algorithms, the overall false positive rate (FPR) and false negative rate (FNR) values of the SGDIDS are calculated by evaluating the whole test dataset. FPR values evaluate the cases where the normal communications are classified as attack scenarios and FNR evaluates the situations where the attacks of DoS, U2R, R2L and Probing are classified as normal communications and the type of attacks which they don't belong to. The records of the FPR and FNR are shown in Table 4.4.

Table 4.4: The Overall FPR/FNR Values of the SGDIDS

Classifier / Values	SVM	CLONALG	AIRS2Parallel
FPR	0.67%	0.7%	1.3%
FNR	2.15%	21.02%	26.32%

From the values of the FPR/FNR in the Table 4.4 it can be concluded that the SVM has stronger classification ability than the AIS algorithms. The main reason for the high FNR

values of the AIS algorithms in the NAN and WAN layers is their failure to classify U2R attacks due to the low occurrence in the training dataset and their classification mechanism.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

As the obsolete electricity and power grid built in last century has brought people various energy problems, the major infrastructure of the power grids should be evolved. The advent and development of the smart grid will introduce more cyber and physical vulnerabilities. Besides taking into account the current cyber attacks in the communication and computer networks as the infrastructure design and deployment, the implementation of the new cyber or physical vulnerabilities and relative approaches of attack detection should also be considered.

The intrusion detection systems have grown considerably in these years, and as the advantage of highly scalable and ease of providing satisfying degradation of service, a large number of distributed intrusion detection systems have been or are being developed and deployed for addressing various needs in current days, but more advanced DIDSs are needed for thwarting the new threats brought by the new equipments, architecture, and networks of the smart grid technology.

In order to improve the cyber security of the smart grid, this work proposed a new architecture, SGDIDS, by utilizing a hierarchical and distributed intrusion detection system. SGDIDS applies a three-layer network composed of home area networks (HAN), neighborhood area networks (NAN), and wide area networks (WAN) that can apply a robust two-way communication network, as well as providing the distributed intrusion detection systems with hierarchical detection abilities to each components in the networks to observe attacks at various levels.

This SGDIDS is applied in the wireless mesh network while also finding the optimal routing for smart grid communications. By using the wireless mesh networks, the scalable, dynamic, and energy-saving communication paths can be provided to create a more efficient communication network for the future smart grid with the more complicated structure. The simulation results illustrate that the optimal communication routes between the commodity pairs can be found simultaneously by using the energy-saving routing searching algorithm. The weights between the constraining pairs are evaluated and the routing of the commodity pairs with lowest weights is selected as the optimal routing.

Security is improved via the classification of intrusion data using the SVM and AIS algorithms. The effectiveness of the SGDIDS for improving security is demonstrated through multiple simulations. Because of the advantages of high accurate classification and the convenient usage, the SVM algorithm and the AIS algorithms are chosen as the classification algorithms in the SGDIDS. By applying these two algorithms on the NSL-KDD dataset, the intrusion detections occur on the related IDS of each node in the network, and sending the testing data to the IDS in the upper layer expressing the hierarchical mechanism. Both SVM and AIS perform well as the detection results have illustrated. All normal and attack communications can be classified by using SVM algorithm.

The accuracy of R2L and Probing detection by AIS algorithms is higher than 90%. The reason of low detection rate of the U2R by AIS is due to the low occurrence of the training dataset and the classification mechanism of the AIS.

5.2 Future Work

Though the two classifiers used in this study have exhibited acceptable or satisfactory performance on the design of the SGDIDS architecture, there is room to further improve their performance.

- The current classifier models will be improved through the inclusion of network packets and attacks that are unique to the smart grid. The parameters of the classifier models will be re-evaluated for better training of the new constructed data.
- It is also possible that some other classifiers may be able to achieve better performance. Thus, one of our future research focuses is to evaluate the effectiveness of some existing classifiers for our problem. In this way, for each attack type the most effective classifier can be determined.
- The optimal combination of different classification techniques and different IDSs can also be developed for achieving a higher overall accuracy for intrusion detection. For instance, the SVM and AIS can be combined as one classifier for different attack types.
- Since the communication of the smart grid will be different from the current computer network, an enhanced dataset derived based on the variations of power system parameters in the presence of cyber attacks will be developed for training and testing classifiers.
- The entire system will also be implemented in a SCADA testbed environment with real-time communications. The intrusions will be emulated in the communications

between servers, and the improved SGDIDS will be implemented to detect the intrusions when they occur.

References

1. U.S. Department of Energy. (2008). *The smart grid: An introduction*. Available: http://www.oe.energy.gov/SmartGridIntroduction.htm,
2. Hauser, C. H., Bakken, D. E., & Bose, A. (2005). *A failure to communicate: next generation communication requirements, technologies, and architecture for the electric power grid*. IEEE Power and Energy Magazine, vol. 3, no. 2, (p. 47-55).
3. Liu, B., Wu, Y., Hu, X., Chen, Y., Zhao, J. (2010). *Analysis of Brazilian Blackout on the November 10, 2009 and Its Revelations*. Southern Power Systems Technology. Vol. 4, no. 1, (p. 23-28).
4. Office of the National Coordinator for Smart Grid Interoperability , *NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 1.0*, Available: http://www.nist.gov/public_affairs/.../smartgrid_interoperability_final.pdf
5. Gungor, V. C. & Lambert, F. C., (2006). *A survey on communication networks for electric system automation*. Comput. Netw, vol. 50, no. 7, (p. 877–897).
6. Cheng, J. & Kunz, T., (2009). *A Survey on Smart Home Networking.* Technical Report SCE-09-10.
7. Qiu, R. C., Chen, Z., Guo, N., Song, Y., Zhang, P., Li, H., & Lai, L. (2010) *Towards a real-time cognitive radio network testbed: architecture, hardware platform, and application to smart grid*. In Proceedings of the fifth IEEE Workshop on Networking Technologies for Software Defined Radio and White Space.

8. Keemink, S. & Roos, B. (2008). *Security analysis of Dutch smart metering systems*. Available: http://staff.science.uva.nl/~delaat/sne-2007-2008/p33/report.pdf
9. Cupp, J. & Beehler, M. (2008). *Implementing Smart Grid Communications*.
10. Daintree Networks. (2007). *What's so good about mesh networks?* Available: http://www.daintree.net/downloads/whitepapers/mesh-networking.pdf
11. The Smart Grid Interoperability Panel – Cyber Security Working Group. *Guidelines for Smart Grid Cyber Security: Vol. 3, Supportive Analyses and References*. Available: http://www.csrc.nist.gov/publications/nistir/ir7628/nistir-7628_vol3.pdf
12. Metke, A. R., & Ekl, R. L. (2010). *Security technology for smart grid networks*. IEEE Transactions on Smart Grid, vol. 1 (p. 99–107)
13. Bennett, C. & Highfill, D. (2008). *Networking AMI Smart Meters*. Energy 2030 Conference, 2008. ENERGY 2008.
14. Oey,M. A., Warnier,M. & Brazier, F.M.T. *Security in Large-Scale Open Distributed Multi-Agent Systems*. Available: http://74.125.155.132/scholar?q=cache:TdSItIbB4SsJ:scholar.google.com/+Security+in+Large-Scale+Open+Distributed+Multi-Agent+Systems&hl=en&as_sdt=0,5
15. Rosenfield, M. G. (2010) *The Smart Grid and Key Research Technical Challenges*. VLSI Technology (VLSIT), 2010 Symposium, 15-17. (p:3-8).
16. McDaniel, P. & Smith, S. W. *Security and Privacy Challenges in the Smart Grid*. Available: http://www.patrickmcdaniel.org/pubs/sp-smartgrid09.pdf

17. Eisenhauer, J., Donnelly, P., Ellis, M., & O'Brien, M. (2006). *Roadmap to Secure Control Systems in the Energy Sector*. [Online]. Available: http://www.controlsystemsroadmap.net/pdfs/roadmap.pdf

18. Liu, Q., Hwang, J.-N., & Liu, C.-C. (2002). *Communication infrastructure for wide area protection of power systems*. Power Syst. Commun. Infrastructures Future, Beijing, China.

19. Ten, C., Manimaran, G., & Liu, C. (2010). *Cybersecurity for Critical Infrastructures: Attack and Defense Modeling*. IEEE Transactions on Systems, Man, and Cybernetics, Part A. (p: 853 – 865).

20. The National Association of Regulatory Utility Commissioners. *The Smart Grid: Frequently Asked Questions for State Commissions*. Available: http://www.naruc.org/Publications/NARUC%20Smart%20Grid%20Factsheet%205_09.pdf

21. National Energy Technology Laboratory. (2009). *A compendium of smart grid technologies*. Office of Electricity Delivery and Energy Reliability, U.S. Department of Energy.

22. Office of Electricity Delivery and Energy Reliability. *A VISION FOR THE SMART GRID*. Available: http://www.netl.doe.gov/smartgrid/referenceshelf/whitepapers/Whitepaper_The%20Modern%20Grid%20Vision_APPROVED_2009_06_18.pdf

23. Zerbst, J., Schaefer, M., & Rinta-Jouppi, I. (2010). *Zone principles as Cyber Security architecture element for Smart Grids*. Innovative Smart Grid Technologies Conference Europe (ISGT Europe), 2010 IEEE PES. (p. 1-8).

24. Cleveland, F. (2008). *Cyber security issues for advanced metering infrastructure (AMI).* 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century. (p. 1–5).
25. McDaniel, P., & McLaughlin, S. (2009). *Security and Privacy Challenges in the Smart Grid.* IEEE Security and Privacy 7(3).
26. Lee, A. & Brewer, T. (2009). *Smart Grid Cyber Security: Strategy and Requirements*. NIST.
27. Dollen, D.V. (2009). *Report to NIST on the smart grid interoperability standards roadmap*. Technical Report SB1341-09-CN-0031, Electric Power Research Institute (EPRI).
28. Goodspeed, T., Highfill, D. R., & Singletary, B. A. (2009). *Low-level Design Vulnerabilities in Wireless Control Systems Hardware.* Proceedings of the SCADA Security Scientific Symposium 2009 (S4). (p. 3-1–3-26).
29. AMI-SEC Task Force and AMI Security Acceleration Project (ASAP). (2009). *AMI Security Implementation Guide V1.01*.
30. Matthew, C. (2009). *Hacking AMI, SANS Process Control and SCADA Security Summit*.
31. Hadley, M.D., McBride, J.B., Edgar, T.W., O'Neil, L.R., & Johnson, J.D. (2007) *Securing Wide Area Measurement Systems*. Prepared for. U.S. Department of Energy.
32. Bobba, R., Heine, E., Khurana, H., & Yardley, T. *Exploring a Tiered Architecture for NASPInet.* 2010 IEEE Conference on Innovative Smart Grid Technologies.
33. International Society of Automation (ISA). (2007). *CS2SAT - Control System Cyber Security Self-Assessment Tool.*
34. BISHOP, M. (2004). *Introduction to Computer Security.* Addison Wesley.
35. Ericsson, G. (2010). *Cyber security and power system communication-Essential parts of a smart grid infrastructure*. In IEEE Transactions of Power Delivery, vol. 25.

36. Davis, C. M., Tate, J. E., Okhravi, H., Grier, C., Overbye, T. J., & Nicol, D. (2006). *SCADA Cyber Security Testbed Development*. Proceedings of the 38th North American Power Symposium, NAPS 2006 Carbondale, IL. (p. 483-488).

37. Fadul, J., Hopkinson, K., Sheffield, C., Moore, J., & Andel, T. (2011). *Trust Management and Security in the Future Communication-Based "Smart" Electric Power Grid*. Proc. 44th Hawaii International Conference on Systems Sciences.

38. Alcaraz, C., Fernandez, G., Roman, R., Balastegui, A., & Lopez, J. (2008).*Secure Management of SCADA Networks. New Trends in Network Management*. Cepis UPGRADE. vol. 9, no. 6. (p 22-28).

39. Tsang, R. *Cyberthreats, Vulnerabilities and Attacks on SCADA Networks*. [Online]. Available: http://gspp.berkeley.edu/iths/Tsang_SCADA%20Attacks.pdf

40. Mukherjee, B., Heberlein, T. L. & Levitt, K. N. (1994) *Network intrusion detection. IEEE Network*. 8(3):26–41.

41. Zamboni, D. (2001). *Using Internal Sensors for Computer Intrusion Detection*. PhD thesis, Purdue University.

42. Balasubramaniyan, J. S., Garcia-Fernandez, J. O., Isacoff, D., Spafford, E., & Zamboni, D. (1998). *An architecture for intrusion detection using autonomous agents*. In Proceedings of the Fourteenth Annual Computer Security Applications Conference, IEEE Computer Society, (p. 13–24).

43. Kumar, S. (1995). *Classification and detection of computer intrusions*. Ph.D. Thesis, Purdue University, West Lafayette, IN 47907

44. Abraham, A., Jain, R., Thomas, J., & Han S.Y. (2007). *D-SCIDS: Distributed soft computing intrusion detection system*. Journal of Network and Computer Applications, vol.30, no.1, (p. 81-98).

45. Kim, G. H., & Spafford, E. H. *The design and implementation of Tripwire: A file system integrity checker*. The 2nd ACM Conference on Computer and Communications Security, Fairfax, Virginia, November 1994. ACM Press. (p. 18–29).

46. Farmer D., & Venema, W. (2000) *Computer forensics analysis class handouts*. online available: http://www.fish.com/forensics/.

47. Crosbie, M. & Spafford, G. (1995) *Active defense of a computer system using autonomous agents*. Technical Report 95-008, COAST Group, Department of Computer Sciences, Purdue University, West Lafayette, Indiana. http://www.cerias.purdue.edu/homes/spaf/tech-reps/9508.ps.

48. Barnett, B. & Dai N. Vu. (1997). *Vulnerability assessment and intrusion detection with dynamic software agents*. Software Technology Conference.

49. Stefan, A. (1999). *Research in intrusion-detection systems: A survey*. TR 98-17, Department of Computer Engineering, Chalmers University of Technology, Sweden, December 1998.

50. Juszczyszyn, K., Nguyen, N. T., Kolaczek, G., Grzech, A., Pieczynska, A., & Katarzyniak, R. (2006). *Agent-based Approach for Distributed Intrusion Detection System Design*. 2006 of the International conference on computational science, Lecture Notes in Computer Science, vol. 3993. Springer Verlag. (p. 224–231).

51. Benattou, M., & Tamine, K. (2005). *Intelligent Agents for Distributed Intrusion Detection System*. Proceedings of World Academy of Science, Engineering and Technology, Volume 6.

52. Chatzigiannakis, V., Androulidakis, G., Grammatikou, M., & Maglaris, B. (2004). *An Architectural Framework for Distributed Intrusion Detection using Smart Agents*. Proceedings of SAM04, Las Vegas.

53. Spafford, E. H. & Zamboni, D. (2000). *Intrusion detection using autonomous agents*. Computer Networks, 34(4). (p. 547-570).

54. Wei, X., Huang, H., & Tian, S. *A Modified RBF Neural Network for Network Anomaly Detection*. ISNN 2006, LNCS 3973. (p. 261 – 266).

55. Wu, S. X., Banzhaf, W. (2010). *The use of computational intelligence in intrusion detection systems: A review.* Applied Soft Computing, journal homepage: www.elsevier.com /locate/asoc. (p. 1–35).

56. Zanero, S. (2005). *Analyzing tcp traffic patterns using self organizing maps*. In: F. Roli, S. Vitulano (Eds.), Proceedings of the 13th International Conference on Image Analysis and Processing – ICIAP 2005, Lecture Notes in Computer Science, vol.3617, Springer, Cagliari, Italy. (p. 83–90).

57. Lichodzijewski, P., Zincir-Heywood, A., Heywood, M.I. (2002). *Dynamic intrusion detection using self-organizing maps.* In: The 14th Annual Canadian Information Technology Security Symposium, Ottawa, Canada.

58. Chimphlee, W., Abdullah, A. H., Sap, M. N. M., Chimphlee, S., & Srinoy, S. (2005). *Unsupervised Clustering methods for Identifying Rare Events in Anomaly Detection*. 6th Internation Enformatika Conference (IEC2005), October 26-28, Budapest, Hungary.

59. Castro, L. N. D. (2006). *Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications*. Chapman & Hall/CRC.

60. Crosbie, M. (1995). *Applying genetic programming to intrusion detection*. In Proceedings of the AAAI 1995 Fall Symposium series.

61. Ramos, V., & Abraham, A. (2005). *ANTIDS: self-organized ant-based clustering model for intrusion detection system*. In: The 4th IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology (WSTST'05), Japan, IEEE Press.

62. Hooper, E. (2010). *Strategic and Intelligent Smart Grid Systems Engineering*. Internet Technology and Secured Transactions (ICITST), 2010 International Conference, London, 8-11. (p. 1-6).

63. Yi, P., Wu, Y., Zou, F., & Liu, N. *A Survey on Security in Wireless Mesh Networks*. IETE Tech. (p.6-14).

64. Garcia-Hernandez, C. F., Ibarguengoytia-Gonzalez, P. H., & Perez-Diaz, J. A. (2007). *Wireless Sensor Networks and Applications: A Survey*. IJCSNS International Journal of Computer Science and Network Security, 7(3). (p.264-273).

65. Aswal, M.S., Rawat, P., & Kumar, T. (2009). *Threats and Vulnerabilities in Wireless Mesh Networks*. International Journal of Recent Trends in Engineering, vol. 2, No. 4, November.

66. Brodsky, J. & McConnell, A. (2009). *Jamming and Interference Induced Denial-of-Service Attacks*. IEEE 802.15.4-Based Wireless Networks, Proceedings of the SCADA Security Scientific Symposium 2009 (S4), January 21–22. (p. 2-1–2-11).

67. Safaric, S. & Malaric, K. (2006). "*ZigBee wireless standard*" 48th International Symposium ELMAR-2006, Zadar, Croatia, 07—09. (p.259-262).

68. Lee, M. J. & al., (2006). *Emerging Standards for Wireless Mesh Technology*. IEEE Wireless Communication.

69. Wei, H., Ganguly, S., Izmailov, A., & Haas, Z. (2005). *Interference-Aware IEEE 802.16 WiMax Mesh Networks*. In Proc. 61st IEEE VTC,

70. Gunn, S. R. (1998). *Support Vector Machines for classification and regression*. Technical report, University of Southampton, Faculty of Engineering, Science and Mathematics; School of Electronics and Computer Science.

71. http://kdd.ccs.uci.edu/databases/kddcup99/task.html

72. Yao, J., Zhao, S., & Fan, L. (2006). *An Enhanced Support Vector Machine Model for Intrusion Detection*. In Proceedings of Rough Sets and Knowledge Technology.

73. Zhang, Z., & Shen, H. (2005). *Application of online-training SVMs for real-time intrusion detection with different considerations*. Computer Communications, vol. 28, Issue 12. (p. 1428-1442).

74. Lee, H., Song, J., & Park, D. (2005). *Intrusion Detection System Based on Multi-class SVM*. RSFDGrC, vol. 2 (p. 511-519).

75. Joachims, T. (1999). *Making large-scale SVM learning practical*. In Advances in Kernel Methods Support Vector Learning, M.I.T. Press.

76. Chang, C. & Lin, C. (2001). *LIBSVM: a library for support vector machines*. [Online] Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

77. Burnet, F.M. (1959). *The Clonal selection theory of acquired immunity*. Vanderbilt Univ. Press, Nashville TN.

78. Castro, D., L. N. & Zuben, F. V. (2000). *The donal selection algorithm with engineering applications*. Proceedings of Genetic and Evolutionary Computation Conference: 36-37.

79. Castro, D., L. N., Zuben, V., & F. J. (2001). *Learning and Optimization Using the Clonal Selection Principle*. IEEE Trans. On Evol. Comp, Special Issue on Artificial Immune Systems.

80. Brownlee, J. (2005). *Clonal selection theory and CLONALG - the donal selection classification algorithm (CSCA)*. Tech. Report 2-01, Centre for Intelligent Systems and Complex Processes, Faculty of Information and Communication Technologies, Swinburne University of Technology, Victoria, Australia.

81. Kim, J., & Bentley, P. (2001).*Towards an Artificial Immune System for Network Intrusion Detection: An Investigation of Clonal Selection with Negative Selection Operator*. Congress on Evolutionary Computation, (p. 1244-1252)

82. Watkins, A. (2001). *AIRS: A Resource Limited Artificial Immune Classifier*. M.S Thesis, Department of Computer Science, Mississippi State University.

83. Watkins, A. & Timmis, J. (2004). *Exploiting parallelism inherent in AIRS, an artificial immune dassifier*. In: Nicosia, ICARIS 2004, LNCS, vol. 3239 (p. 427-438).

84. Manoa Innovation Center. *The Case for Using Wireless Mesh Technology in Smart Grid Communication Networks*. [Online]. Available: http://www.concentris-systems.com/downloads/SmartGridWMAdvantages.pdf

85. Sikora, A. & Groza, V. F. (2005). *Coexistence of IEEE 802.15.4 with other systems in the 2.4 GHz-ISM-Band*. IEEE Instrumentation & Measurement Technology Conference, Ottawa. (p. 1786-1791).

86. Wu, J., Wang, C., Wang, J., & Chen, S. (2006). *Dynamic hierarchical distributed intrusion detection system based on multi-agent system*. In WI-IATW '06: Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, Washington, DC, USA, IEEE Computer Society. (p. 89–93).

87. Kodialam, M. & Nandagopal, T. (2005). *Characterizing the Capacity Region in Multi-Radio Multi-Channel Wireless Mesh Networks*. In Proc. ACM MobiCom.

88. Cortes, C. & Vapnik, V. (1995). *Support-vector networks in Machine Learning*. vol. 20. (p. 273–297). [Online]. Available: http://www.springerlink.com/index/K238JX04HM87J80G.pdf

89. Kim, D. & Park, J. *Network Based Intrusion Detection with Support Vector Machines*. ICOIN 2003, LNCS 2662

90. Gunn, S. R. (1998). *Support vector machines for dassification and regression*. Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science,

Tech. Rep. [Online]. Available: http://pubs.rsc.org/en/Content/ArticlePDF/2010/AN/B918972F/2009-12-23

91. Lanaridis, A., Karakasis, V., Stafylopatis, A. (2008). *Clonal selection-based neural classifier*. In: 2008 8th international conference on hybrid intelligent systems (HIS). (p. 655–660).
92. Kim, J., & Bentley, P. (2001). *Towards an Artificial Immune System for Network Intrusion Detection: An Investigation of Clonal Selection with Negative Selection Operator.* Congress on Evolutionary Computation. (p. 1244-1252).
93. Khaled, A., Abdul-kader, H.M., & Nabil, A. (2010). *Artificial Immune Clonal Selection Classification Algorithms for Classifying Malware and Benign Processes Using API Call Sequences.* IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.4. (p. 31-39).
94. Catal. C. & Diri, B. (2009). *Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem.* Information Sciences, Vol. 179, Issue 8. (p. 1040-1058).
95. Brownlee, J. (2005). *Artificial immune recognition system (airs) – a review and analysis. Technical report*. Victoria, Australia: Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology.
96. Catal, C., & Diri, B. (2007). *Application and benchmarking of artificial immune systems to classify fault-prone modules for software development projects*. IADIS International Conference Applied Computing. (p. 347-354).
97. http://wekaclassalgos.sourceforge.net/

98. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. (2009). *A Detailed Analysis of the KDD CUP 99 Data Set*. IEEE International Conference on Computational Intelligence for Security and defense applications. (p. 53-58).

99. Benferhat, S., Sedki, K., Tabia, K. (2007) *Preprocessing rough network traffic for intrusion detection purposes* IADIS International Telecommunications, Networks and Systems. (p.105-109).

100. Berthier, R., Sanders, W., & Khurana, H. (2010). I*ntrusion Detection for Adcanced Metering Infrastructures: Requirements and Architectural Directions*. In First IEEE International Conference on Smart Grid Communications (SmartGridComm). (p. 350–355).

101. Castro, D., L. N. & Timmis, J. (2002). *An Artificial Immune Network for Multimodal Function Optimization*. Proceedings of the IEEE Congress on Evolutionary Computation (CEC'02), Vol. 1, May, Hawaii. (p. 699-674).

Appendix A

Screenshots of Weka Platform for AIS Classifiers

A.1 Preprocess Window (User-Defined)

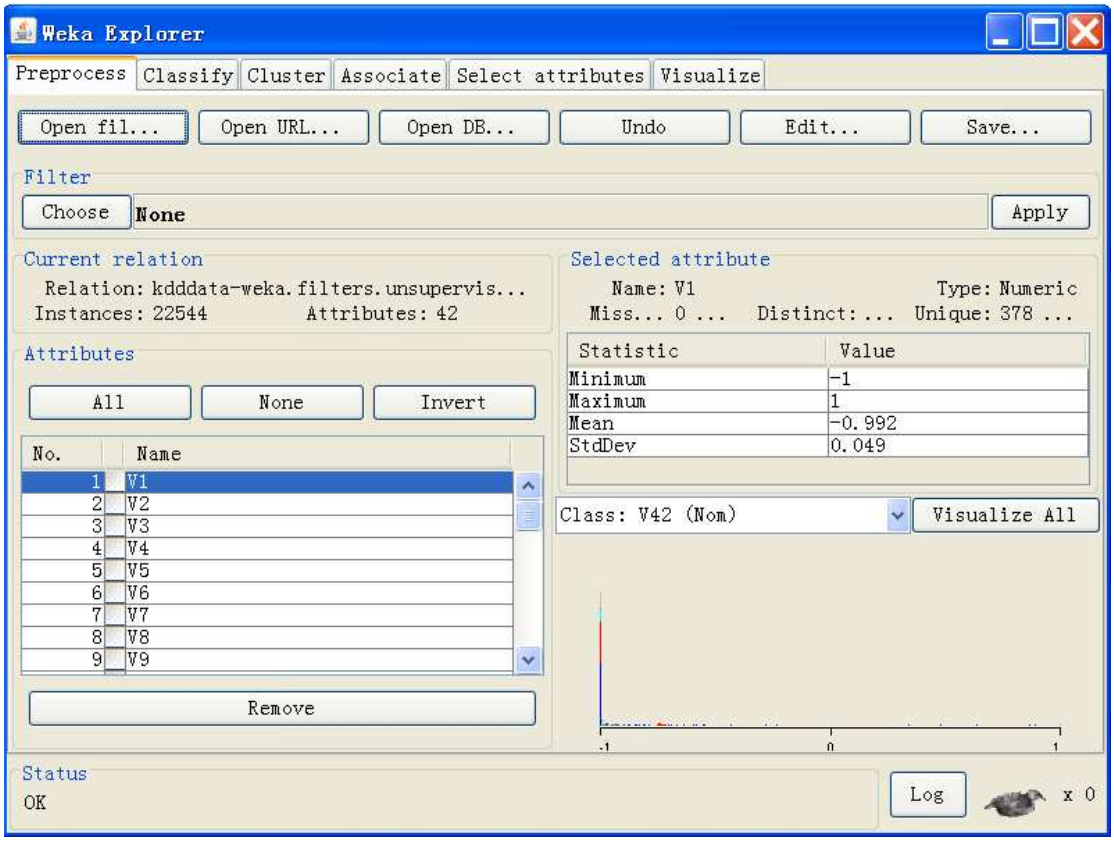

Figure A.1: Screenshot of user-defined selection of data preprocess window.

A.2 Classifier Selection Window (User-Defined)

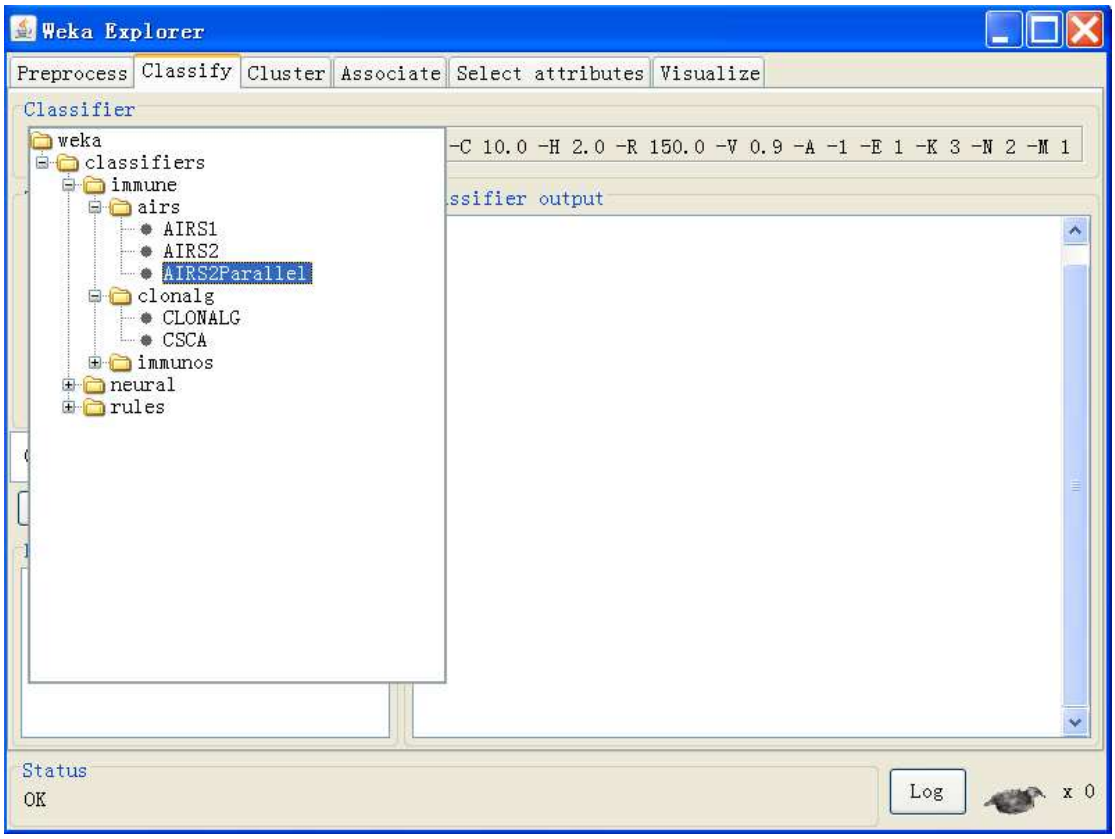

Figure A.2: Screenshot of user-defined selection of classifier algorithms window.

A.3 Classification Using CLONALG Model

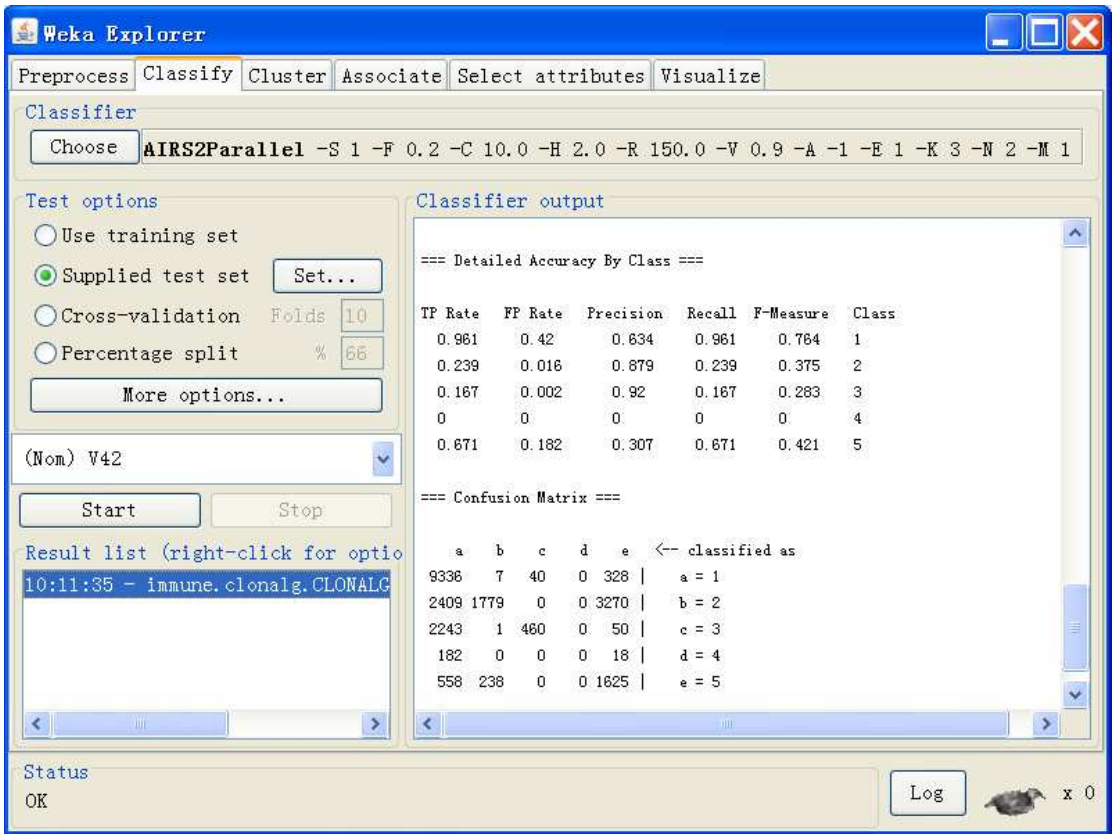

Figure A.3: Screenshot of classification result of the CLONALG model.

A.4 Classification Using AIRSParallel Model

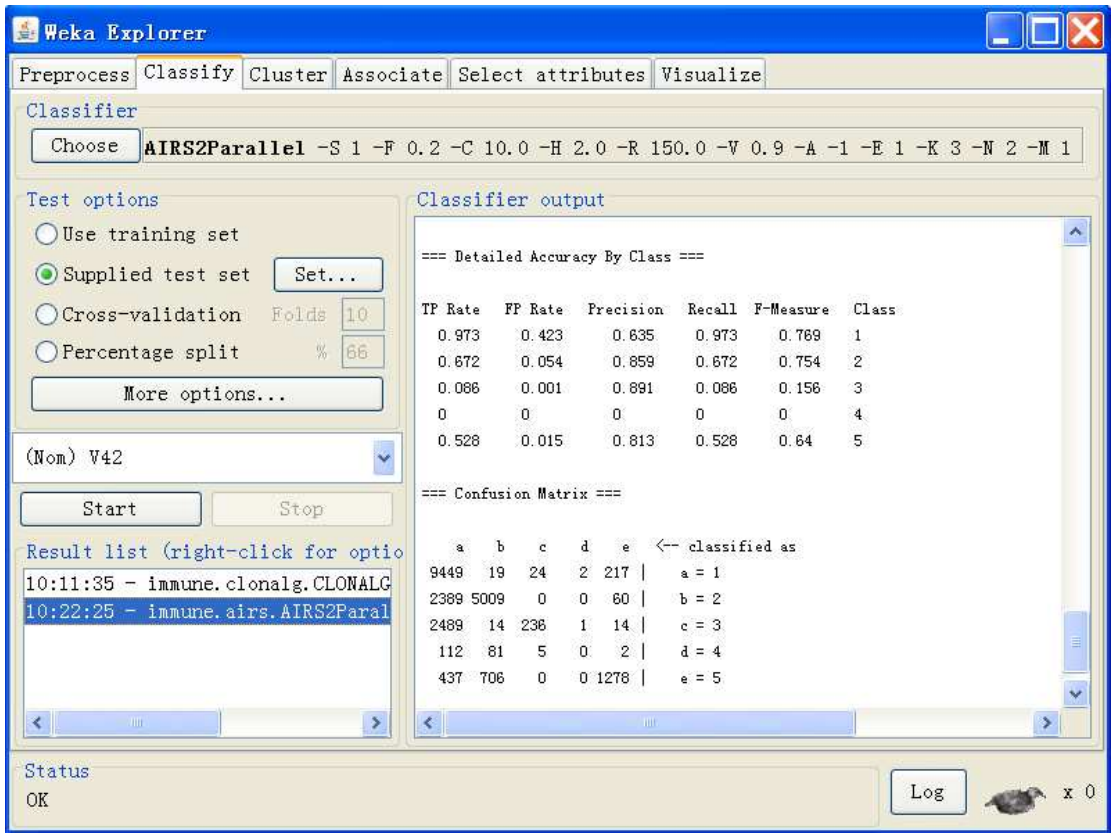

Figure A.4: Screenshot of classification result of the AIRSParallel model.

Appendix B

Source Code for Primal-Dual Optimal Algorithm and Classifier Algorithms (Matlab)

B.1 Source Code for Primal-Dual Routing Optimization Algorithm

```
epslo = 0.2;                                                % parameter setting
Flowie = zeros(sizex,1);
weightj = ones(sizex,1)*10^(-3);
totalweight = sum(weightj);
b = 0;
while(totalweight < 1)
    for q = 1:numcommodity
        r = R(q);
        while (r>0)
            sq = commodity(q,1);
            dq = commodity(q,2);                            % get the starting and eding node
            [dist,path,pred] = graphshortestpath(UG,sq,dq);
            pathrecord{q}=path;
            if (size(path)==0)
                fprintf('node %f no connection.',sq);
                break;
            end
            nodenum = size(path,2);
            flowpath = zeros(nodenum-1,1);
            for i = 1:(nodenum-1)
                pathtemp = path(1,i:i+1);
                if (pathtemp(1)<pathtemp(2))            % start to find the optimal route
                    pathnum1 = find(secon == pathtemp(1));
                    pathnum2 = find(revercon == pathtemp(2));
                    pathnum=intersect(pathnum1,pathnum2);
                    flowpath(i,1)=X(pathnum,1)+idsflow;
                    if(pathtemp(2)~=size(C,1))

plot([C(pathtemp(1),1)+13,C(pathtemp(2),1)-12],[C(pathtemp(1),2)+13,C(pathtemp(2),2)-12],'Color',[1
```

```
-linkcolortemp(commodity(q,1),1)/round(numcommodity*1.5),0.5,linkcolortemp(commodity(q,1),1)/r
ound(numcommodity*1.5)],'LineWidth',2);
    % connect two nearest nodes and illustrate the path
                        end
                        if(pathtemp(2)==size(C,1))

plot([C(pathtemp(1),1)+37,C(pathtemp(2),1)-18],[C(pathtemp(1),2)+5,C(pathtemp(2),2)-23],'Color',[1-
linkcolortemp(commodity(q,1),1)/round(numcommodity*1.5),0.5,linkcolortemp(commodity(q,1),1)/r
ound(numcommodity*1.5)],'LineWidth',2);
                        end
                    else
                        pathnum1 = find(secon == pathtemp(2));
                        pathnum2 = find(revercon == pathtemp(1));
                        pathnum=intersect(pathnum1,pathnum2);
                        flowpath(i,1)=X(pathnum,1)+idsflow;
                        if(pathtemp(1)~=size(C,1))

plot([C(pathtemp(2),1)+13,C(pathtemp(1),1)-12],[C(pathtemp(2),2)+13,C(pathtemp(1),2)-12],'Color',[1
-linkcolortemp(commodity(q,2),1)/round(numcommodity*1.5),0.5,linkcolortemp(commodity(q,2),1)/
round(numcommodity*1.5)],'LineWidth',2);
                        end
                        if(pathtemp(1)==size(C,1))

plot([C(pathtemp(2),1)+37,C(pathtemp(1),1)-18],[C(pathtemp(2),2)+5,C(pathtemp(1),2)-23],'Color',[1-
linkcolortemp(commodity(q,2),1)/round(numcommodity*1.5),0.5,linkcolortemp(commodity(q,2),1)/r
ound(numcommodity*1.5)],'LineWidth',2);
                        end
                    end
                end
                u = min(flowpath);                          % get the shortest routing
                sigma = min(u,r);
                r = r-sigma;
                for i = 1:(nodenum-1)
                    pathtemp = path(1,i:i+1);
                    if (pathtemp(1)<pathtemp(2))
                        pathnum1 = find(secon == pathtemp(1));
                        pathnum2 = find(revercon == pathtemp(2));
                        pathnum=intersect(pathnum1,pathnum2);
                        Flowie(pathnum) = Flowie(pathnum)+ sigma;
                    % information flow accumulation within the capacity
                        while (Flowie(pathnum)>Capacity(pathnum))
                            Flowie(pathnum) = Flowie(pathnum)-sigma;
                            newpathtemp = newpath;
                            if (newpathtemp(1)<newpathtemp(2))
                                newpathnum1 = find(secon == newpathtemp(1));
                                newpathnum2 = find(revercon == newpathtemp(2));
                                pathnum=intersect(newpathnum1,newpathnum2);
                                flowpath=X(pathnum,1)+idsflow;
                                u = min(flowpath);
                                sigma = min(u,r);
                                Flowie(pathnum) = Flowie(pathnum)+ sigma;
                            else
                                pathnum1 = find(secon == pathtemp(2));
                                pathnum2 = find(revercon == pathtemp(1));
                                pathnum=intersect(pathnum1,pathnum2);
                                flowpath =X(pathnum,1)+idsflow;    % this is the new X
                                u = min(flowpath);
                                sigma = min(u,r);
```

```
                        Flowie(pathnum) = Flowie(pathnum)+ sigma;
                    end
                end
                if (size(flowpath,1) == 1)
                    weightj(pathnum) = weightj(pathnum)*(1+epslo*sigma/flowpath);
                    break;
                end
                weightj(pathnum) = weightj(pathnum)*(1+epslo*sigma/flowpath(i,1));
            else
                pathnum1 = find(secon == pathtemp(2));
                pathnum2 = find(revercon == pathtemp(1));
                pathnum=intersect(pathnum1,pathnum2);
                Flowie(pathnum) = Flowie(pathnum)+ sigma;
                while (Flowie(pathnum)>Capacity(pathnum))
                    Flowie(pathnum) = Flowie(pathnum)-sigma;
                newpathtemp = newpath;
                if (newpathtemp(1)<newpathtemp(2))
                    newpathnum1 = find(secon == newpathtemp(1));
                    newpathnum2 = find(revercon == newpathtemp(2));
                    pathnum=intersect(newpathnum1,newpathnum2);
                    flowpath = X(pathnum,1)+idsflow;
                    u = min(flowpath);
                    sigma = min(u,r);
                    Flowie(pathnum) = Flowie(pathnum)+ sigma;
                else
                    pathnum1 = find(secon == pathtemp(2));
                    pathnum2 = find(revercon == pathtemp(1));
                    pathnum=intersect(pathnum1,pathnum2);
                    flowpath =X(pathnum,1)+idsflow;
                    u = min(flowpath);
                    sigma = min(u,r);
                    Flowie(pathnum) = Flowie(pathnum)+ sigma;
                end
                end
                if (size(flowpath,1) == 1)
                    weightj(pathnum) = weightj(pathnum)*(1+epslo*sigma/flowpath);
                  % increase the weight of the nodes
                    break;
                end
                weightj(pathnum) = weightj(pathnum)*(1+epslo*sigma/flowpath(i,1));
            end
        end
    end
    b = b+1;
    totalweight = totalweight + sum(weightj);
end
rol = max(Flowie);
bestramda = b/rol;
```

B.2 Source Code for AIS Classifier Selection Algorithm

```
case 1
```

```
            type = number*10 + 1;               %use the CLONALG model
            ModelName=['kdd' num2str(type) '.model'];
        case 2
            type = number*10 + 2;               % use the AIRSParallel model
            ModelName=['kdd' num2str(type) '.model'];
end
VData = [attack_inst,attack_label];
WekaString = char('weka.classifiers.trees.J48');      % open Weka
S=size(VData,2);
for i=1:S-1
        attributeName{i}=['V' num2str(i)];
        attributeType{i}='numeric';
end
attributeName{S}=['V' num2str(S)];
attributeType{S}='{1,2,3,4,5}';
arffWrite('KDDtest.arff','Verification',attributeName,attributeType,VData);       %    write the data
type %into the arff file
ActualPath=cd;
WekaPath=which('wekaclassalgos.jar');
ModelOut='ModelOut.txt';
Temp=['java -cp "' WekaPath '" ' WekaString ' -l "' ActualPath '\' ModelName '" -T "' ActualPath
'\KDDtest.arff"    -p 5 > "' ModelOut '"'];                                    % get the test results
dos(Temp);
mm = 4;
testtemp = zeros(1,4);
fid =fopen('ModelOut.txt');
for i = 1:mm
        headerSt = fscanf(fid,'%s/n');
        testtemp(i) = str2double(headerSt);
```

```
end
predictvalue = testtemp(2);
realvalue =    testtemp(4);
if (predictvalue == realvalue)
      accuracy = 1;
else
      accuracy = 0;
end
predict_label = predictvalue;
```