# CHAPTER 2
# LITERATURE REVIEW

## 2.1 BACKGROUND

This chapter discusses and explains the concept and theory of both BIS and AIS and later presents the relationship between these two by mapping them with each other. The chapter is divided in few subsections. After discussing on the AIS and BIS, this chapter highlights current researches on spam and presents related studies that focus upon detection and classification of spam messages. Figure 2.1 shows the methodological framework of the research.

From the Figure 2.1, the author should able to understand and identify the concept of BIS and AIS through literature review (i.e. view existing previous research and journal) and manage to map how these both concepts related to each other. Then these concepts will be applied in spam management by proposing IMSM to manage the problem of spam messages in mobile phone. Three phases involved in IMSM; detection, classification and severity determination.

In the detection phase, two types of AIS algorithms were used; Danger Theory and Negative Selection to detect spam and ham messages. Their performances were compared to identify which one is better and three proposed features were introduced to enhance the

performance of Danger theory. Several experiments were conducted and results were discussed and analysed.

In the classification phase, Immune Network Theory and Clonal Selection were used to classify the spam messages into several group of spam. HICNA was introduced for classification and three phases involved (i.e. cluster using common keywords, cluster using uncommon keywords and lastly using expert judgement).

Severity determination is to determine the level of danger for each spam messages according to the type of spam. Three level of danger were used; low, medium and high. This phase is beyond the scope of this research and it anticipated that the severity determination will be covered and studied in PhD level.

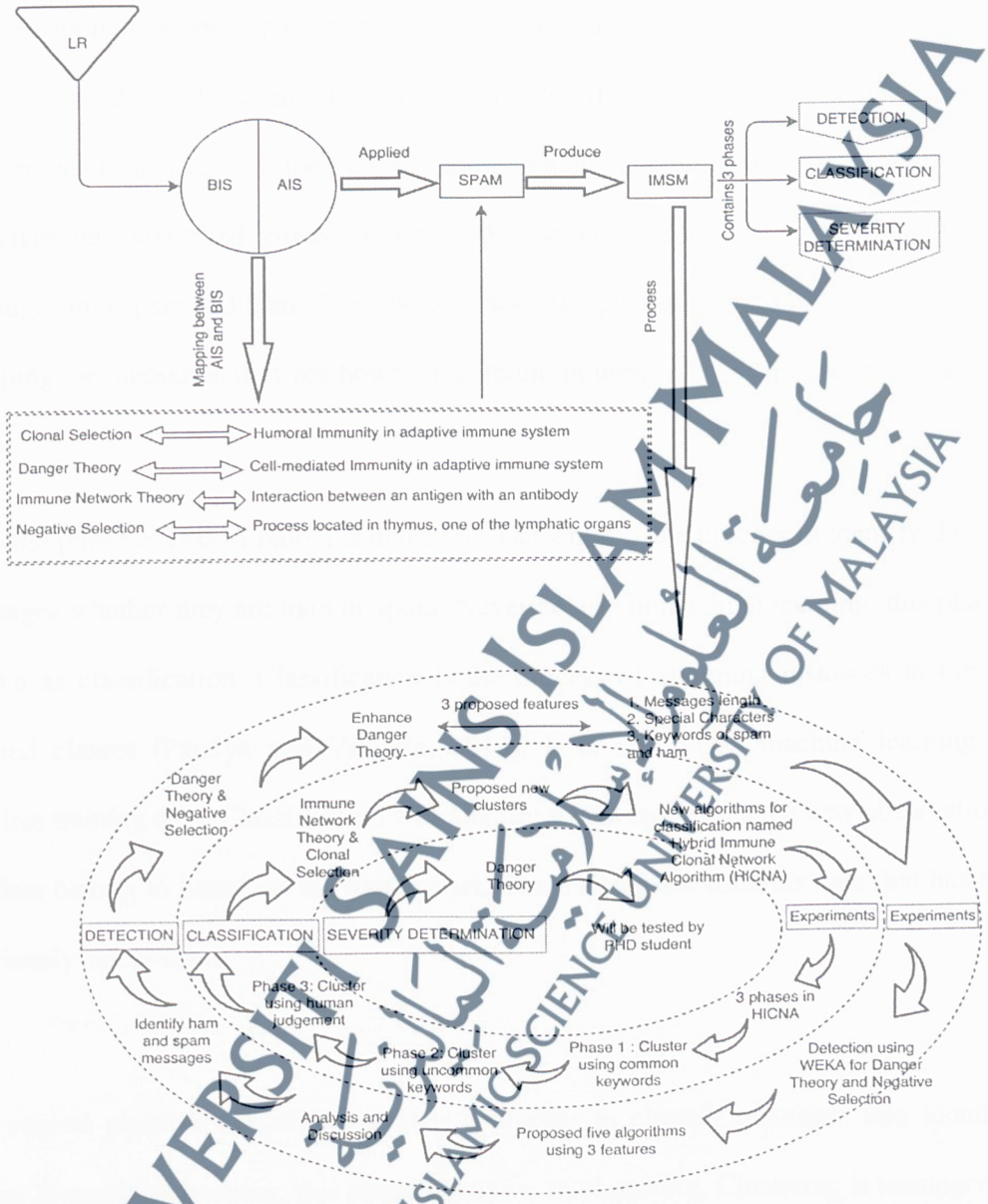


**Figure 2. 1**: The methodological framework of the research

## 2.2 DEFINITION OF TERMS

This research proposes a spam management model named Integrated Mobile Spam Model (IMSM) and three phases involved; detection, classification and severity determination. This research focuses on the second phase which is classification to classify spam messages into identified group. However, the detection is needed to identify the SMS messages into spam and ham. Even though these two phases seem to be similar in term of grouping the messages, they are however different in terms of concept and meaning.

The first phase of IMSM model is detection. Detection is the process to identify the SMS messages whether they are ham or spam. Nevertheless, in machine learning, this phase is known as classification. Classification is the process of assigning instances to the pre-defined classes (Pandya and Virparia, 2013). It is supervised machine learning and requires training data. Classification will identify which categories the new observation or the data belong to based on the training set. The training set contains data that has been previously categorised.

The second phase is classification. It is a process to classify messages into identified group. In machine learning, this phase is known as clustering. Clustering is unsupervised learning and does not require training data. Clustering can be used to find groups of documents with similar content (Chaudhari and Parikh, 2012). It tries to group a set of objects and find whether there is a relationship between objects in groups.

Previous published papers (Christina et al, 2010; Chaurasia and Pal, 2014) related to detection process show that there is a similar concept of clustering with the concept of detection. However in this thesis, we can conclude that **classification is the process of identifying groups of documents that have been defined using training data while clustering will identify groups that have similar characteristics and try to categorize the unlabelled data into groups.**

## 2.3 IMMUNE SYSTEM

The immune system is a system in the human body to protect against diseases and foreign microorganisms such as viruses and bacteria (Purves et al., 2010). It is a complex system that can protect our body and maintain a healthy body. The multifunction of the immune system inspired researchers to develop algorithms to be used for the optimization and computer security areas.

### 2.3.1 BIOLOGICAL IMMUNE SYSTEMS

The human body is made up of cells similar to cells of animals and plants. A group of cells with the same structure and function will produce tissues like muscle. An organ (i.e. lungs) is made from a group of different tissues thus from a group of organs will produce the systems (i.e. excretory system and reproduce system). The unique and varieties of structures and components in our body produce complex system such as immune system. All living things are endowed with complex and various immune systems according to

their characteristics to protect themselves from being attacked from enemies. For example, the rose flower has thorns in its brunch and squid produces black liquid when they are in danger.

The immune system is important because it can prevent and protect the human body from foreign invaders that can cause infections to the body. This system is an amazing system because of its ability to recognize millions different types of enemies and quickly responds to the dangerous components (i.e. virus and bacteria) in our body. On top of that, it can also remember previous enemies that attacked the body and give immunity when similar attack occurs for the second time.

Generally, there are three types of blood in our body; platelets (thrombocytes), red blood cells (erythrocytes) and white blood cells (leukocytes). All of these blood are developed from hematopoietic stem cells and formed in the bone marrow through the process of haematopoiesis; the formation of blood cellular components (About.com Biology, 13 January 2014). Platelets help the blood clotting process by gathering at the site of injury. Red blood cells carry oxygen from the lung to the rest of the body and then return carbon dioxide from the body to lungs while white blood cells are important in the immune system. Figure 2.2 shows the formation of blood cells in human body.

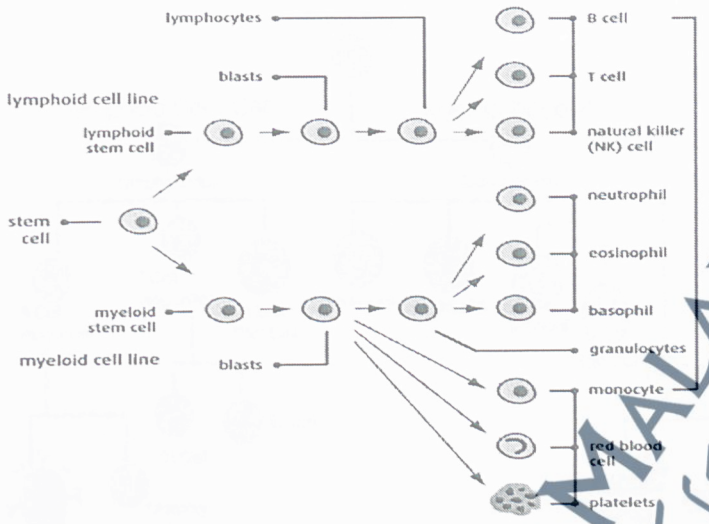


**Figure 2.2**: Formation of blood cells

(Purves et al., 2010)

The immune system is developed and originated from white blood cells in the bone marrow. White blood cells act as a protection and destroy foreign antigens such as bacteria and viruses in the body. There are two categories of white blood cells (Figure 2.3) which are granulocytes and lymphocytes. Granulocytes are cells that contain small particles in their cytoplasm and are important to kill the foreign substances. Five types of granulocytes white blood cells are neutrophil, eosinophil, basophil, mast cell and monocyte (dendritic cell and macrophage) and all of them play a role in the innate immune system (i.e. nonspecific response). Table 2.1 shows the function for each granulocyte white blood cells and it can be said that most of them have an ability to engulf antigens.

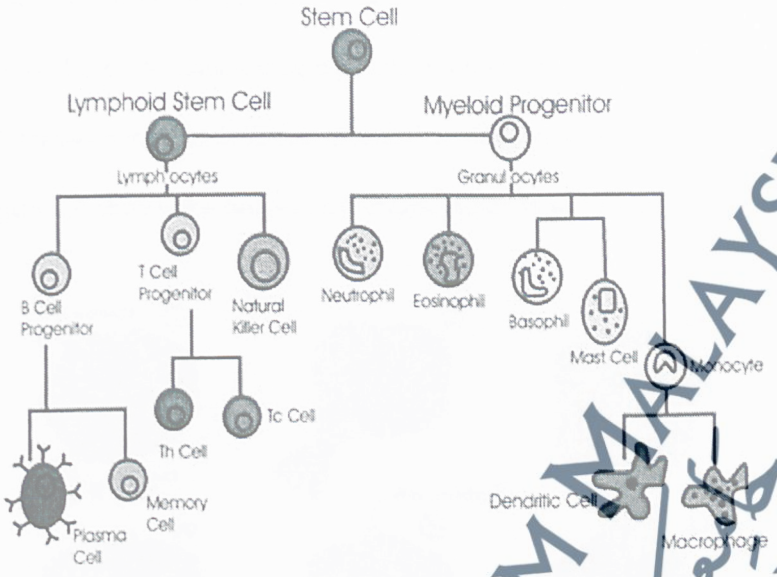


**Figure 2. 3**: White Blood Cells classification

(Todar's Online Textbook of Bacteriology, 5 March 2014)

**Table 2. 1:** Functions of Granulocytes White Blood Cells

| Types | Function |
|---|---|
| Basophils | Releases histamine and involved in inflammatory reactions in the body, especially those related to allergies and asthma. |
| Eosinophils | Kills antibody-coated parasites. |
| Neutrophils | Phagocyte (i.e. they can ingest other cells or particles) and the first immune cells to arrive at a site of infection. |
| Mast cells | Release histamine when damaged. |
| Monocytes | Develop into macrophages and dendritic cell. |
| Macrophages | Engulf and digest microorganism; activate T cells. |
| Dendritic cells | Present antigens to T cells and B cells. |

Granulocytes white blood cells act as a phagocyte. A phagocyte is a process of phagocytosis (i.e. process is used by cells to engulf and ingest harmful foreign particles) and this process is a major mechanism in the innate immune system. In phagocyte, white

blood cells such as macrophage will engulf and surround the unwanted particles or foreign microbes. Then, the macrophage breaks it down by mixing it with enzymes stored in sacs called lysosomes. The leftover material is then pushed out of the cells as waste materials. Figure 2.4 shows the process of phagocytosis in our body.

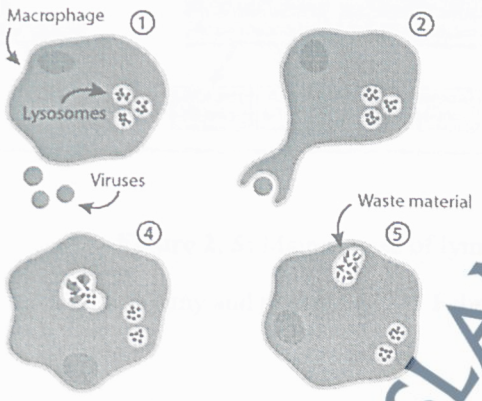


**Figure 2. 4**: Process of phagocytosis by Macrophage (ASU, 4 March 2014)

Another category of white blood cells is known as lymphocytes as it activates in lymphatic tissues. The lymphatic system is a network of lymph capillaries and large vessels that empties into the circulatory system and plays an active role in defending the body from pathogens (LiveScience, 15 February 2014). Bone marrow, thymus, spleen, lymph nodes are the example of lymphatic organs and important in the immune system. Lymphocytes consist of B-cells, T-cells and Natural Killer cells that help to defend our body from internal or external attacks. Figure 2.5 shows the main classes of lymphocytes.

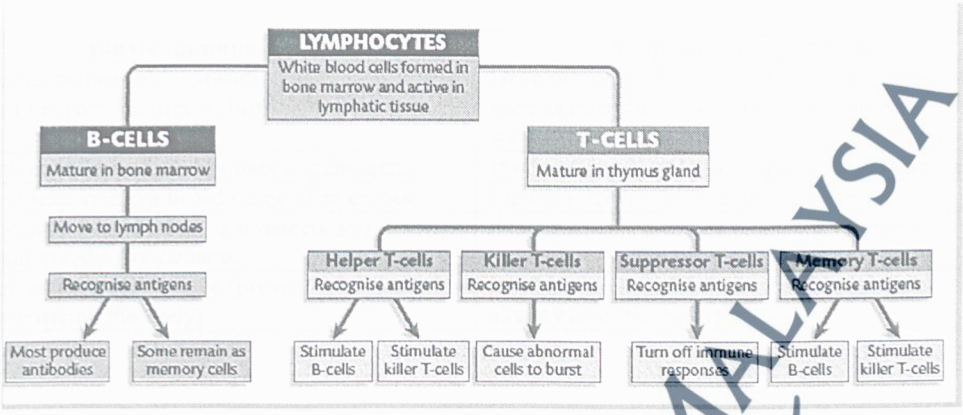


**Figure 2. 5:** Main classes of lymphocytes

(Anatomy and physiology, 23 February 2014)

Two main classes of lymphocytes are called B cells and T cells and they are used in the adaptive immune system (i.e. specific response). B cells produce antibodies and T cells regulate the production of antibodies by B cells. Another type of lymphocytes is natural killer cell but this cell functions in innate immune response to identify virus infected cells like tumour cells and kills them by attacking the cell membrane, causing the cell to burst.

The body utilizes many different types of immunities to protect itself from infections that seemingly endless. This defence may be external that prevents pathogens (i.e. virus and bacteria) from entering the body and also the internal defence that fights pathogens that have already entered the body. Two types of defences are innate immune system and adaptive immune system (InnerBody, 2 March 2014). Table 2.2 explains the difference between these two immune systems (Purves et al., 2010; Todar, 2012).

**Table 2. 2:** Innate and Adaptive Immune System

| Innate Immune System | Adaptive Immune System |
|---|---|
| Presents before any exposure to pathogens and is effective from the time of birth. | Develops only after exposure to inducing agents such as microbes, toxins or others foreign substances. |
| Involves nonspecific responses to pathogens. | Involves a very specific response to pathogens. |
| Rapid response to a broad range of microbes. | Slower response to specific microbes. |
| Unchanging mechanism that detects and destroys certain invading organisms. | Responds to previously unknown foreign cells and remembers when the attacking occurs again. |
| Involves external defence (prevents pathogens from entering the body). | Involves internal defence (attacks pathogens that already enter the body). |

The innate immune system is the nonspecific defence mechanism and it naturally presents in our body. Granulocytes white blood cells take a role in this system by preventing pathogens from attacking our body. This system consists of two types of defences; the first line defence and second line defence. In the first line of defence, the mechanism focuses on the prevention of external body like skin and mucous membrane. Skin produces sweat glands that contain oily and acidic liquids to inhibit pathogens while mucus with its viscous fluid can trap pathogen. In the second line of defence, the process involves phagocytosis to kill pathogens.

The adaptive immune system is the third line of defence and it involves a specific defence mechanism. Humoral immunity and cell-mediated immunity are two classes in adaptive immunity. Humoral immunity involves defence from extracellular bacteria and is divided into two types (i.e. plasma cells and memory cells) while cell-mediated immunity defence intracellular bacteria are divided into four types of T cells. Figure 2.6 presents the theoretical summary of BIS in the human body.

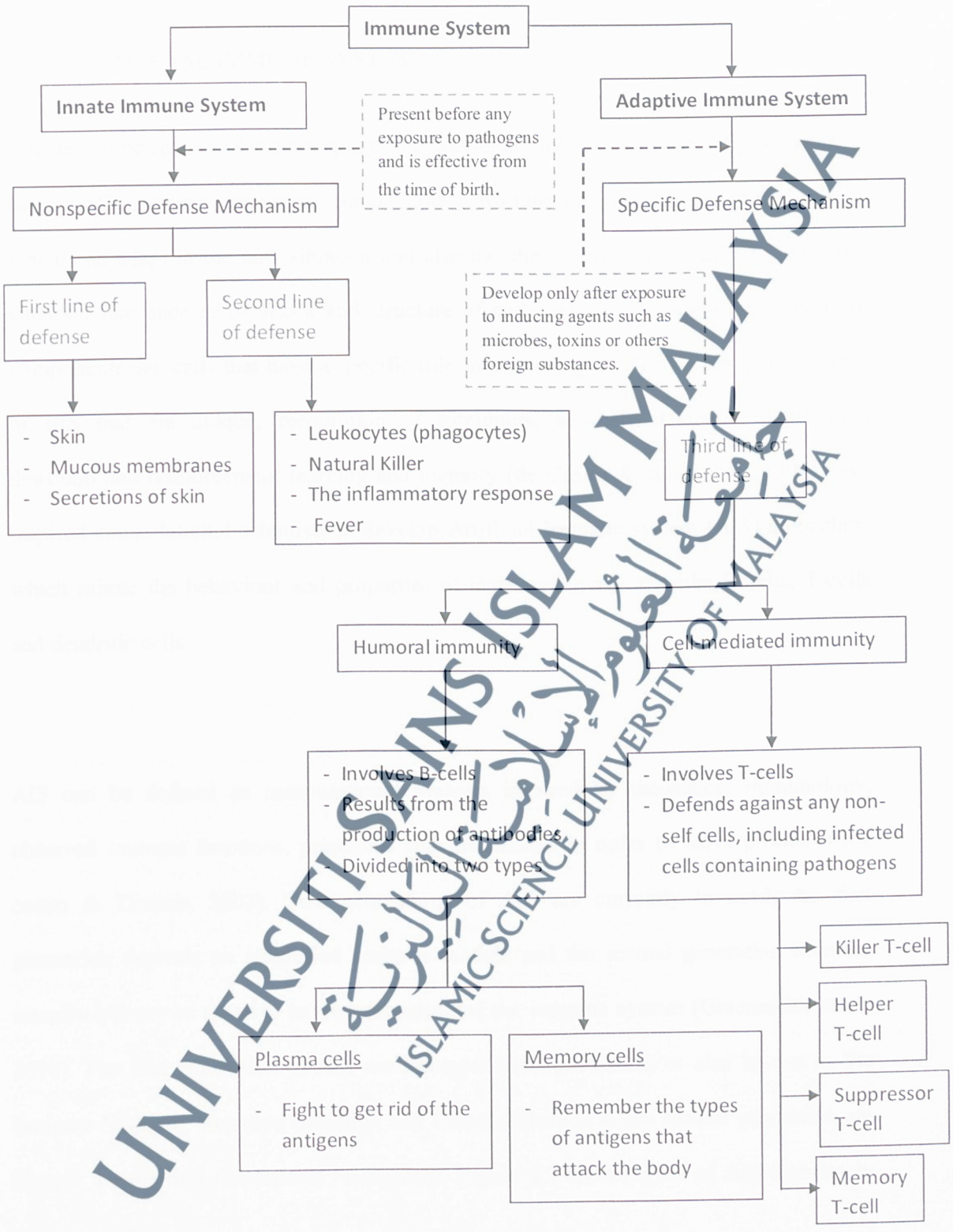


**Figure 2. 6**: Theoretical summary of BIS

### 2.3.2 ARTIFICIAL IMMUNE SYSTEM

The design perspective for developing computational tools inspired naturally is termed as biologically inspired computing (Nanda, 2009). An immune system is a complex system which can adapt to the new situation and also has the ability to remember the previous situation like patterns of attack and structure of pathogen. It also contains a variety of components and cells that have a specific role in protecting the body. The characteristics of BIS that are unique, recognition of foreigners, anomaly detection, distribution detection and reinforcement learning and memory (de Castro & von Zuben, 1999) have inspired computational scientists to develop Artificial Immune system (AIS) algorithms which mimic the behaviour and properties of immunological cells like B-cells, T-cells and dendritic cells.

AIS can be defined as computational systems inspired by theoretical immunology, observed immune functions, principles and mechanism in order to solve problems (de castro & Timmis, 2002). Two generations of AIS are currently in used; the first generation depends on simplified immune models and the second generation involves interdisciplinary to develop an understanding of the immune system (Greensmith et al., 2010). The first generation of AIS are Idiotypic (network-based) or also known as the Immune Network, Negative Selection and Clonal Selection while second generation are Danger Theory and Conceptual Framework. Figure 2.7 shows types of AIS theories in each generation.

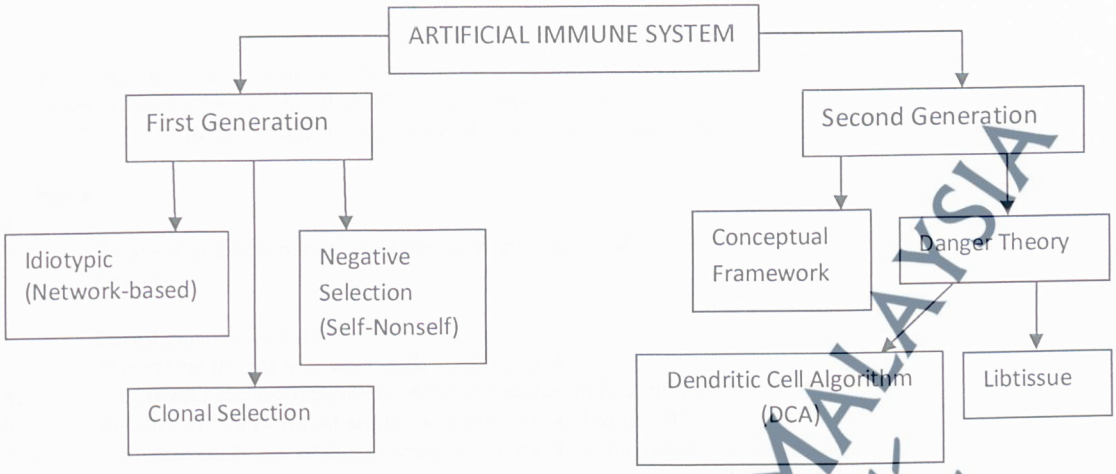


**Figure 2.7**: AIS Theories

a. Immune Network Theory

In the early 1970s, a famous immunologist working in Switzerland introduced a new theory to further understand the immune system, Jerne (1974) proposed that cells and molecules of the immune system do not only recognize foreign substances but also recognize, respond to and are regulated by each other. This ideology is known as the Idiotypic Network Theory, or more simply the Immune Network Theory. The theory states that the interaction or binding of lymphocytes is not only with the foreign molecules but can also make a connection between each of them because they have variable (v) regions. Thus, the immune system has a network with the components connected to each other by V-V interactions. Figure 2.8 shows the algorithm for the Immune Network Theory.

input: S = set of patterns to be recognized, *nt* network affinity threshold, *ct* clonal pool threshold, *h* number of highest affinity clones, *a* number of new antibodies to introduce
output: N = set of memory detectors capable of classifying unseen patterns

```
1. begin
2.
3.      Create an initial random set of network antibodies, N
4.      repeat
5.
6.      Forall patterns in S do
7.       Determine the affinity with each antibody in N
8.         Generate clones of a subset of the antibodies in N with the highest affinity. The
9.       number of clones for an antibody is proportional to its affinity
10.        Mutate attributes of these clones to the set A , a and place h number of the highest
11.     affinity clones into a clonal memory set, C
12.      Eliminate all elements of C whose affinity with the antigen is less than a predefined threshold ct
13.        Determine the affinity amongst all the antibodies in C and eliminate those
14.     antibodies whose affinity with each other is less than the threshold ct
15.        Incorporate the remaining clones of C into N
16.     end
17.
18.     Determine the affinity between each pair of antibodies in N and eliminate all
19.     antibodies whose affinity is less than the threshold nt
20.      Introduce a random number of randomly generated antibodies and place into N
21.
22.     end until a stopping criteria has been met
23. end
```

**Figure 2. 8:** The algorithm for Immune Network Theory

(AISWEB, 5 February 2014)

b. Clonal Selection

When antibodies on a B-cell bind with an antigen, the B-cell becomes activated and begins to evolve. The new B-cells clones are produced from the parent B-cell but then undergo the somatic hypermutation (Berek & Ziegner,1993) and produce antibodies that are specific to the invading antigen. The Clonal Selection is proposed by Burnet in 1959 and his theory described the mechanism by which immune cells can generate memory to a specific antigen. The idea is that only those cells capable of recognizing an antigen

stimulus will proliferate, thus being selected against those that do not (Timmis & Knight, 2002). In this theory, cells that can recognize an antigen will activate and undergo mitosis process and generate more antibodies and memory cell to remember the structure and component of the antigen. Figure 2.9 shows the algorithm for Clonal Selection.

```
input: S = set of patterns to be recognized, n the number of worst elements to select for removal
output: M = set of memory detectors capable of classifying unseen patterns

1. Begin
2.
3. Create an initial random set of antibodies, A
4.
5.      Forall patterns in S do
6.          Determine the affinity with each antibody in A
7.          Generate clones of a subset of the antibodies in A with the highest affinity.
8.          The number of clones for an antibody is proportional to its affinity
9.          Mutate attributes of these clones to the set A ,
10.         place a copy of the highest affinity antibodies in A into the memory set, M
11. Replace the n lowest affinity antibodies in A with new randomly generated antibodies
12.end
13.    end
```

**Figure 2. 9**: The algorithm for Clonal Selection.

(AISWEB , 5 February 2014)

c. Negative Selection

The immune system is complete because it has the ability to recognize all antigens including the previous antigens and respond to them. The purpose of negative selection is to provide tolerance for self-cells. This theory deals with the ability of immune system to detect unknown antigens while not reacting to the self-cells (Nanda, 2009). The immune system needs to be able to distinguish between the molecules of our own cells (self) and foreign molecules (non-self) in order to function properly. Negative Selection occurs in thymus, one of the lymphatic organs and involves T-cells. During Negative Selection, T

cells interact with the thymic dendritic cell. T cells with high-affinity interaction are eliminated through apoptosis (to avoid autoimmunity), and those with intermediate affinity survive. The first Negative Selection algorithm was proposed by Forrest et al., (1994) to detect data manipulation caused by a virus in a computer system. Figure 2.10 shows the algorithm of negative selection.

input: $S_{seen}$ = set of seen known self elements
output: D = set of generated detectors

1. **begin**
2.
3. **repeat**
4. Randomly generate potential detectors and place them in a set *P*
5. Determine the affinity of each member of *P* with each member of the self set $S_{seen}$
6. If at least one element in *S* recognises a detector in *P* according to a recognition threshold,
7. then the detector is rejected, otherwise it is added to the set of available detectors *D*
8. **until** Stopping criteria has been met
9. **end**

**Figure 2. 10**: The algorithm for Negative Selection.

(AISWEB, 5 February 2014)

d. Danger Theory

This theory was found by Matzinger (Matzinger, 2002; Matzinger, 1998; Matzinger, 1994). The possibility that discrimination occurs in the immune system which is it does not care about self and non-self that led Matzinger to come up with this theory. The main idea in this theory is that the immune system does not respond to non-self but to danger. The danger is measured by damage cells showed by distress signals that are sent out when cells die in unnatural death (cell stress or lytic cell death, as opposed to programmed cell

death, or apoptosis). She pointed out that there must be discrimination that happens and goes beyond the self-non-self distinction. For instance:

- There is no immune reaction to the foreign bacteria in the gut or to the food we eat although both are foreign entities.
- Conversely, some auto-reactive processes are useful, for example against self-molecules expressed by stressed cells.
- The definition of self is problematic – realistically, the self is confined to the subset actually seen by the lymphocytes during maturation.
- The human body changes over its lifetime and thus self-changes as well. Therefore, the question arises whether defence against non-self learned early in life might be auto-reactive later.
- Autoimmune diseases and certain types of tumours are fought by the immune system (both attacks against self) and successful transplant (no attack against non-self).

```
input: S= set of data items to be labeled safe or dangerous
output: D = set of data items labeled safe or dangerous

1. Begin
2.          Create an initial population of dendritic cells (DCs), D
3.          Create a set to contain migrated DCs, M
4. Forall data items in S do
5.          Create a set of DCs randomly selected from D , P
6.          Forall DCs in P do
7.                    Add data item to DCs collected list
8.                    Update danger, PAMP and safe signal concentrations
9.                    Update concentrations of output cytokines
10.                   Migrate the DC from D to M and create a new DC in Dif concentration of
11.co-stimulatory molecules is above a threshold
12.         end
13.end
14.
15.    Forall DCs in M do
16.         Set DC to be semi-mature if output concentration of semi-mature cytokines is
17.greater than mature cytokines,
18.         otherwise set as mature
19.end
20.
21.         Forall data items in S do
22.         Calculate number of times data item is presented by a mature DC and a semi-mature DC
23.          Label data item a safe if presented by more than semi-mature DCs than mature DC's,
24.otherwise label as dangerous
25.Add data item to labeled set M
26.
27.end
28.   end
```

**Figure 2. 11**: The algorithm for Danger Theory

(AISWEB, 5 February 2014)

e. Conceptual Framework

Stepney et al., (2004) suggested that bio-inspired algorithms are best developed and analysed in the context of a multidisciplinary conceptual framework that provides sophisticated biological models and well-founded analytical principles and they introduce

a framework (the Conceptual Framework) for the successful development of AIS. Four stages are identified as key:

- Observation and experiments: provide a (partial and noisy) view of the complex biological system and probe using practical experimentation.
- Models: validate simplifying abstract representations of the biology.
- Algorithms: Computational systems are developed, implemented, and studied theoretically using the abstract models as a blueprint.
- Applications: the developed algorithms are applied to specific problems, with feedback to the algorithm for refinement.

The lack of rigour in the metaphors used to inspire AIS lead to the development of the conceptual framework. A framework for constructing algorithms is certainly necessary in principle, since it clearly defines the role each discipline must play, i.e. observation by immunologists, modelling by mathematicians, algorithm development by computer scientists and application testing by engineers (Greensmith et al., 2010).

## 2.4 THE RELATIONSHIP BETWEEN AIS AND BIS

Figures 2.12 to 2.14 highlight the mapping of AIS algorithm to their original BIS and Table 2.3 presents the summary of AIS and BIS.

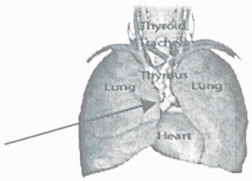


**Figure 2. 12**: Negative Selections in Thymus

The thymus is one of the lymphatic organs. It is a place for T lymphocytes to mature and also the place for the process of negative selection and positive selection. In positive selection, the cells that have no interaction with self-MHC (Major Histocompatibility Complex) expressed by thymic epithelial cells are destroyed while in the negative selection, T-cells with high affinity (i.e. the strength) interaction with thymic dendritic cells are eliminated and those with intermediate affinity survive.


**Figure 2. 13**: The Immune Network Theory in an Antibody and Antigen

The Immune network theory is inspired from the interaction of antibodies with antigens or with themselves. An antibody binds with an antigen to destroy the antigen or also known as a foreign microorganism. The interaction of binding between these two is with the help of epitope and paratope. The epitope is also known as an antigenic determinant

and it is the part of an antigen that is recognized by the immune system. Paratope is the site in an antibody that binds to an epitope of an antigen. Besides, antibodies can also recognize and bind with each other to get more strength of connection to destroy antigens.

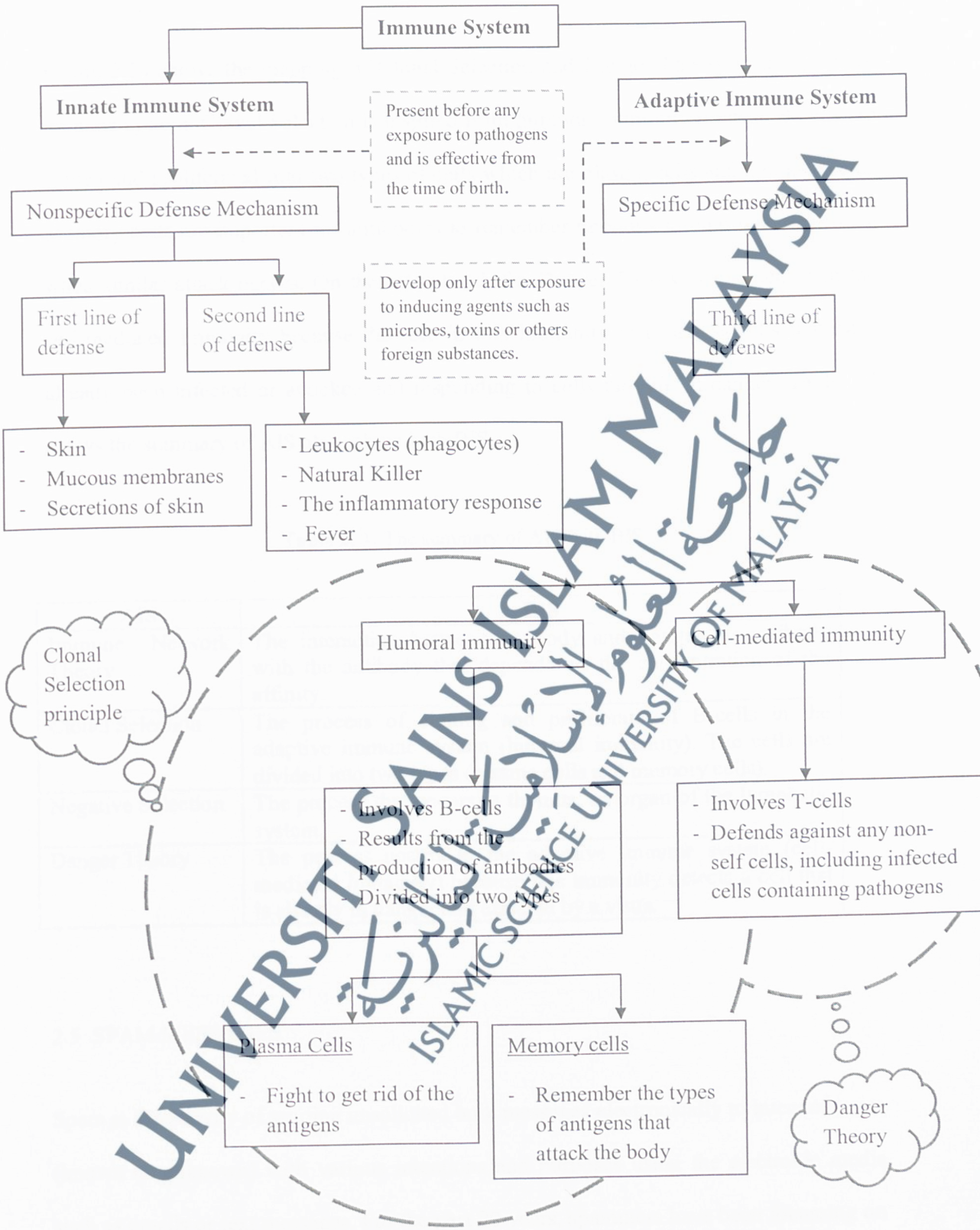




**Figure 2. 14**: Mapping the Danger Theory and Clonal Selection with BIS

Figure 2.14 shows the mapping of Clonal Selection and Danger Theory in the adaptive immune system. Clonal Selection is inspired from humoral immunity because B-cells are cloned and proliferated into two types of cells which are plasma cells and memory cells. Memory cells are important for antibodies to remember previous foreign microorganism when similar attack occurs. On the other hand, the Danger Theory is inspired from the cell-mediated immunity because the role in this immunity is to attack cells that have already been infected or attacked and responding to cells that are in danger. Table 2.3 shows the summary of AIS algorithms with BIS.

**Table 2. 3:** The summary of AIS with BIS

| AIS | BIS |
|---|---|
| Immune Network Theory | The interaction between antibody and antigen or antibody with the antibody that depends on the concentration of the affinity. |
| Clonal Selection | The process of cloning and proliferate of B-cells in the adaptive immune system (humoral immunity). The cells are divided into two types (plasma cells and memory cells). |
| Negative Selection | The process that occurs in thymus, an organ of the lymphatic system. |
| Danger Theory | The process occurs in the adaptive immune system (cell-mediated immunity) because this immunity detects a cell that is already in danger and affected by a virus. |

## 2.5 SPAM MESSAGES

Spam is the activity of sending unsolicited bulk messages electronically to intended users (known or unknown) with various intentions and purposes using the electronic media such as emails or text messages. For the past 20 years, spammers have been focussing on

emails as their usage was very important at that time. However today, with the enhancement of mobile technology and the sophistication related to the new invention of smartphones and tablets, it lead spammers to change their target to mobile users. Generally, authorities and researchers have taken various countermeasures in order to reduce and control spam that is rising, but the rate of spam continues to rise.

The issue of mobile phone spam has not received much attention from researchers as they are more familiar with email spam. However, the SMS spam that influences users directly as they look at every message received is one of the bigger effects for mobile users compared to email. The short messages consist of few words composed of abbreviations and idioms are another challenge in filtering SMS spam. Table 2.4 shows the differences between email and SMS in terms of length and presentation (Sharma & Paul, 2013).

**Table 2. 4:** Differences between Email and SMS

| Feature | Email | SMS |
|---|---|---|
| LENGTH | Unlimited. | 160 in English characters or 70 Arabic and Chinese. |
| REPRESENTATION | Texts, images, attachment, etc. | Only text. |

### 2.5.1 RESEARCH IN SPAM

To date, many techniques are being developed and investigated by researchers to overcome and reduce spam problem (Iqbah et al., 2016). Pour et al., (2012) discussed three techniques for email spam detection; namely *list-based*, *statistical algorithm* and *IP-*

*based*. The *list-based* technique is classified into three categories, namely Blacklist, Whitelist and Greylist. Blacklist blocks the IP address based on complaints from recipient (Ramachandran et al., 2006), while Greylist rejects mail from unknown sources on the theory that real mailers will retry the mail and spammers will not (Levin, 2005). The Whitelist is a technique where any received email from an address that is not stored in the email contact list is rejected and considered uncertain. The *statistical algorithm* can be categorised into the content-based method and rule-based method. The content-based method is commonly used and it filters the content of mail body and headers. It uses machine learning which needs to be trained (Nosrati & Pour, 2011). For example, Chakraborty and Mondal (2012) applied different decision tree classifiers to filter spam mail while Amayri and Bouguila (2010) used Support Vector Machine (SVM) for spam filtering. Besides, Afzal and Mehmood (2016) also used machine learning to filter spam in tweeter. Rule based method works through certain rules and these rules will decide to pass or block the email (Nosrati & Pour, 2011). Reverse lookup is an example of a method in the *IP-based* technique. It is a method of resolving an IP address into a domain name (Rouse, 2007).

Although different approaches and techniques have been introduced, spam messages are still flooding in emails and SMS. This is believed due to the swift adoption of new techniques by spammers and the inflexibility of spam filters to adapt the changes. The next subsections highlight related works or studies related to spam classification and detection in order to further understand 'state-of-the-art' and define the gap for developing spam management model.

a) Detection/Filtering

Recent studies use machine learning for detecting process either in education, health and banking (Chaurasia & Pal, 2014; Amin & Habib, 2015; Ilic et al., 2016; Khare et al., 2016). In machine learning, this phase is known as classification. Classification is one of the techniques in supervised learning (i.e. has training set and testing set). The role of classification is to predefine class and know in which class a new object belongs to (Rahman & Afroz, 2013). There are many classifier algorithms; such as Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN) and Feed Forward Neural Networks (FFNN). Each of these classifiers has their own role and definition so their performances are different based on data used. There are also different data mining tools used for classification process and comparing the accuracy of classifiers algorithm as discussed by (Rahman & Afroz, 2013). They compared various classification techniques using WEKA, TANAGRA and MATLAB. Results show that TANAGRA machine learning tool is the best compared to WEKA and MATLAB. Table 2.5 summarises the existing researches related to the performance of classifiers using different datasets.

**Table 2. 5:** Existing researches related to the performance of classifiers

| Author (S) | Year | Title | Aim | Technique/ Classifiers Used | Results |
|---|---|---|---|---|---|
| Abdullah H. Wahbeh and Mohammed Al-Kabi | 2012 | Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text | Compare the performance of three text classification techniques using a set of Arabic text documents and the document set falls into four categories. | Support Vector Machine (SVM ), Naïve Bayes (NB) and C4.5. | NB classifier achieves the highest accuracy followed by SVM and c4.5.<br>The SVM requires the lowest amount of time to build the model needed to classify, followed by NB and c4.5. |
| Ahmad LG, Eshlaghy AT, Poorebrahmi A, Ebrahimi M, and Razavi AR | 2013 | Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence | Compare the performance of three well-known algorithms through sensitivity, specificity and accuracy using breast cancer recurrence in patients who were doing follow-up for 2 years. | Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN). | The SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The predicted accuracy for DT model is the lowest of all. |
| Shabia Shabir Khan and Mushtaq Ahmed Peer | 2013 | Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques | Evaluate various classification data mining algorithms using WEKA and concentrate on the values of certain important evaluation measures. | a. Rule-based ZeroR and oneR<br>b. Bayes Theorem based Naïve Bayes<br>c. Neural Network based<br>d. Multi-Layer Perceptron<br>e. Decision Tree based J48 and Random Forest | Naïve Bayes classifier is the best classifier for credit dataset. However, this may not be same for all the datasets. |
| Sohil Pandya and Paresh V. Virparia | 2013 | Comparing the Application of Various Algorithms of Classification Techniques of Data Mining in an Indian University to Uncover Hidden Patterns | Examine and investigate various methods of classification to identify the best fit methods among them for the University domain. | Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (k-NN) Feed Forward neural Networks (FFNN), Support Vector Machine (SVM). | FFNN and k-NN are more accurate with their average of ROC is similar, fair and the highest compared to other classifiers. In addition, the error rate in FFNN is the lowest compared to others. |
| Vikas Chaurasia, and Saurabh Pal | 2014 | A Novel Approach for Breast Cancer Detection using Data Mining Techniques | Analyse the breast cancer data with the aim of developing accurate prediction models for breast cancer using data mining techniques and investigate the performance of different classification techniques. | Sequential Minimal Optimization (SMO), IBK (k-Nearest Neighbour), BF Tree. | SMO has higher prediction accuracy than IBK and BF Tree methods. |
| Md. Nurul Amin, and Md. Ahsan habib | 2015 | Comparison of Different Classification Techniques using WEKA for Haematological Data | Show the comparison of different classification algorithms using WEKA and find out which algorithm is the most suitable for users working on haematological data. | Decision Tree (DT), Naïve Bayes (NB) and Neural Network (NN). | NB classifier has the lowest average error compared to DT and NN. Besides, NB classifier has the potential to significantly improve the conventional classification method for being used in medical or in general bioinformatics field. |

With respect to spam messages classification, Shahi and Yadav (2013) investigated mobile SMS spam filtering for Nepali text using Naïve Bayesian and Support Vector Machine (SVM). They identified performance and efficiency of these two types of filters (i.e. Naïve Bayesian and SVM) in the problem of Nepali SMS spam and reported 87.15% accuracy for SVM and 97.74% accuracy for Naïve Bayesian. Besides, Chakraborty and Mondal (2012) used three types of decision tree classification techniques (i.e. Naïve Bayes Tress (NBT), C4.5 Decision Tree (J48) and Logistic Model Tree (LMT)) to analyse their performance for spam mail filtration using WEKA. Results showed that LMT was more accurate classifiers and J48 took minimum time than other algorithms. In 2011, evolutionary learning classifiers (i.e. FuzzyadaBoost, Genetic clASSifier system, eXtended classifier system and sUpervised classifier system (UCS)) were used by Junaid and Faroq for mobile spam filtering. Results showed that UCS achieves more than 89% detection rate and 0% false alarm rate than other classifiers. Both Chabra et al., (2010) and Joe and Shim in 2010 used Support Vector Machine (SVM) for filtering spam messages in email and SMS and results indicated that this classifier can give better results in detecting spam and ham. Christina et al., (2010) employed supervised machine learning techniques (i.e. C 4.5 Decision Tree classifier, Multilayer Perceptron and Naïve Bayes) to filter the email spam messages and the model was built by training with known spam emails and legitimate emails using 10-fold cross validation. Results found that Multilayer Perceptron classifier outperformed other classifiers and the false positive rate was also very low compared to other algorithms.

Several methods were also introduced for filtering spam messages without using classifier algorithms. Pour et al., (2012) introduced new and efficient approach to prevent spam emails from being transferred by shifting the location of the filtering system from the recipient mail server to the sender mail server so that the time required to detect and avoid spam is minimum. Meanwhile, Nosrati and Pour in 2011 proposed a new algorithm known as the concept drift detection with three different levels (i.e. control level, warning level and alarm level) and another algorithm called the Dynamic Concept Drift Detection (DCDD). The proposed algorithm manages to detect sudden concept changes of spam attack with more accuracy. Bing et al., (2010) proposed a three-way decision approach to filter email spam based on the Bayesian spam filter (i.e. accept, reject and further-exam) and the new approach reduces the error rate of classifying a legitimate email to spam, and provides a better spam precision and weighted accuracy.

Generally, SMS spam can be detected by examining and reviewing message contents (i.e. *content features*) or the way messages were sent (i.e. *non-content features*). Sohn et al., (2009) proposed a method of using the stylistic information to the content-based mobile spam filtering. They focused on the way the SMS was written (i.e. stylistic aspect) and four features of stylistic were used - the length of messages, function word frequencies, part-of-speech n-grams and special characters. Tan et al., (2012) identified features of SMS spam based on word and character grams, and alphanumeric and non-alphanumeric characters. They also tested a number of statistical features such as message length, and proportion of uppercase letters and punctuation. Another study was by Xu et al., (2012) where they focused on the non-content features such as statistical, temporal and network

features. Their study showed that the temporal and network features were more effective as compared to the statistical features. Uysal et al., (2013) investigated the impact of several feature extraction and feature selection approaches on filtering of SMS messages in two different languages, namely Turkish and English. There are six lists of structural features extracted from SMS messages given and the lists are message length, the number of terms, uppercase character ratio, non-alphanumeric character ratio, numeric character ratio and presence of URL. Recently, Mujtaba and Yasin (2014) used four features for detecting SMS spam - size of the message, the existence of frequently occurring monograms in the messages, existence of frequently occurring diagram in the message and message classes. These features were implemented and trained in machine learning algorithm for better accuracy. Another study was done (Mosquera et al.,2014) where they analysed the effectiveness of machine learning filters based on linguistic and behavioural patterns in order to detect short text spam. Two different filtering systems have been proposed for message filtering and abusive sender identification. The obtained results show the validity of the proposed solution by enhancing the baseline approaches.

There are also published papers that use AIS for detecting spam. In the work by Victor et al., (2013) and Idris (2012), the Negative Selection algorithm was used to distinguish the characteristics of self and non-self messages from trained dataset. The results from both experiments confirmed that the proposed model is able to establish a better true positive on the unknown spam. Zhu and Tan (2011) proposed a technique using the Danger Theory for spam detection. They proposed a Danger Theory (DT) based learning (DTL) model for combining classifiers and this model mimics the mechanism of DT. Results

suggested that the theory can be used in spam detection. Later, Idris and Muhammad (2012) improved email classification method based on Artificial Immune System to reduce false positive and create spam detector. Results showed that the process is very effective in reducing the false rate (i.e. false positive and false negative). Mahmoud and Mahfouz (2012) proposed a mobile agent system for detecting SMS spam based on AIS and the performance of this system was compared with the Naïve Bayesian algorithm in terms of true positive, true negative, false positive, false negative, detection rate, false positive rate and overall accuracy. Results suggested that the performance of the proposed system is better than Naïve Bayesian algorithm.

b) Classification/Clustering

Classification or mostly known as clustering is the process of grouping a collection of objects into classes of similar objects (Patel et al., 2014). Clustering is an unsupervised learning because it tries to find hidden structure in unlabelled data. It splits the data into groups of similar objects (Babur et al., 2015). For example, grouping patient records with similar symptoms without knowing what the symptoms indicate. The cluster is an ordered list of objects which have some common characteristics (TRIPOD, 8 July 2014). Thus clustering is a division of data into groups with similar objects and each group (i.e. cluster) consists of similar objects between themselves and dissimilar objects with other groups. There are many methods available for clustering process, with the most used are known as the Partitioning method, Hierarchical method, Frequent itemset-based method, Concept-based method, Density-based method, Constraint-based method, Seeded method,

Ontology based method, Grid-based method and Model-based method (Prabha et al., 2014).

According to Chen et al. (2010), document clustering is divided into two major subcategories; hard clustering and soft clustering. Hard clustering (disjoint) assigns each document to exactly one cluster while soft clustering (overlapping) allows each document to appear in multiple clusters (Shah & Mahajan, 2012). Soft clustering is divided into partitioning, hierarchical and frequent itemset-based as shown in Figure 2.15.

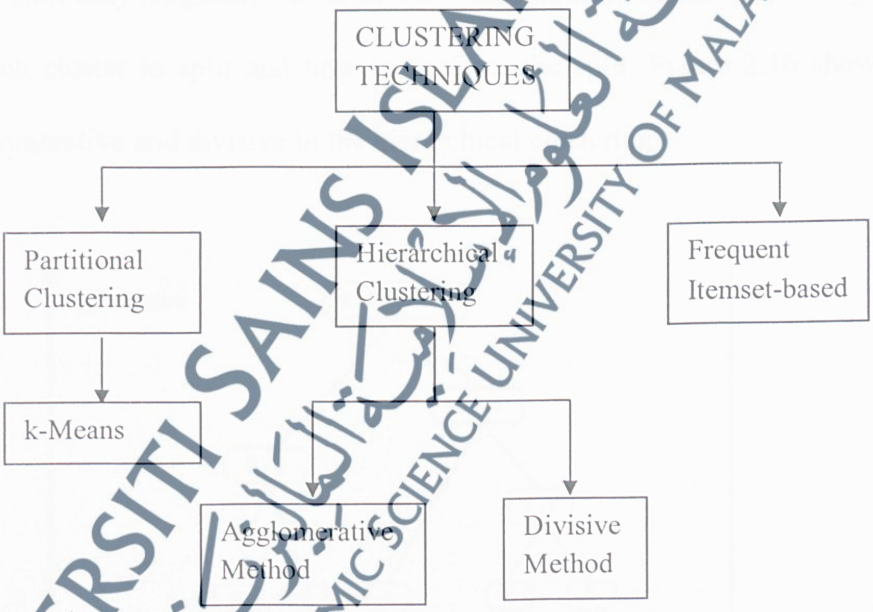


**Figure 2. 15**: Clustering techniques

The hierarchical technique builds a tree-based hierarchical taxonomy (dendrogram) from a set of documents which leaf nodes represent the subset of a document collection. Two basic approaches in this technique are agglomerative and divisive.

Agglomerative or also known as bottom-up start with the points as individual clusters and each step merge with the most similar or closest pair of clusters. This approach requires cluster similarity or distance between each cluster. For example, we have five clusters; A, B, C, D and E. If cluster A has the closest distance with cluster E, they will merge and become one cluster, AE. It is the same if cluster AE has similarity with cluster B, they will merge and become into one big cluster that is AEB.

Divisive or top down starts with one single cluster. It divides and splits into sub-cluster and smaller one until only singleton cluster of individual point remains. In this stage, we must know which cluster to split and how to perform the split. Figure 2.16 shows an example of agglomerative and divisive in the hierarchical clustering.

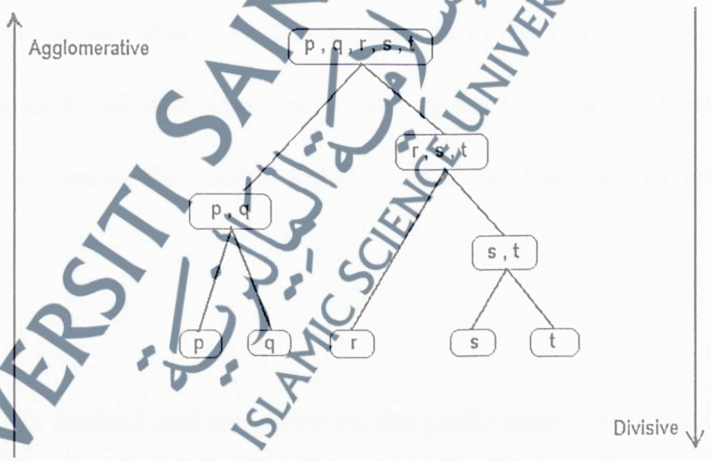


**Figure 2. 16**: Agglomerative and Divisive clustering

(FrontlineSolvers, 2014)

Partitional clustering is another technique used for clustering and this method usually requires the specification of the number of clusters. It constructs k partition of data and evaluates them by some criterion. There are a number of partitional techniques used such as k-Means clustering, Partitioning Around Medoids (PAM), Self-Organizing Maps (SOM) and model-based clustering. k-Means clustering is the most widely used because it is one of the simplest unsupervised learning algorithms that solve the well-known clustering problems. k-Means is based on the idea that a centre point can represent a cluster.

Another type of clustering technique is frequent itemset-based. This method uses frequent itemset generated by the association rule mining to cluster the documents (Shah & Mahajan, 2012). The advantages of using this method are each cluster can be labelled by the obtained frequent itemset shared by the documents in the same cluster. Besides it can reduce the dimensionality of term features efficiency for very large datasets. A frequent itemset is an asset of words that occur together in some fractions of a document in a cluster.

The previous research studied and reviewed on the performance of clustering algorithms using data mining tool which is WEKA. WEKA is a collection of machine learning algorithms for data mining tasks and it contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Table 2.6 shows existing researches related to clustering process using WEKA.

**Table 2. 6:** Existing studies related to the performance of clustering algorithms

| Author | Year | Title | Aim | Clustering Algorithms | Results |
|---|---|---|---|---|---|
| Pallavi and Sunila Godara | 2011 | A Comparative Performance Analysis of Clustering Algorithms | Analyse and compare the performance of three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. | Hierarchical Method, k-Means Method and Farthest First Method. | k-Means algorithm performs well without inserting the principle component analysis filter as compared to the Hierarchical and Farthest First clustering algorithms. |
| Narendra Sharma, Aman Bajpai, and Ratnesh Litoriya | 2012 | Comparison the Various Clustering Algorithms of WEKA Tools | Make comparison of different clustering algorithms of WEKA and find out which algorithms will be the most suitable for the users. | COBWEB clustering algorithm, DBSCAN clustering algorithm, EM clustering algorithm, Farthest First clustering algorithm, OPTICS clustering algorithm and k-Means clustering algorithm. | k-Means clustering algorithm is the simplest algorithm as compared to others. |
| Bharat Chaudhari, and Manan Parikh | 2012 | A Comparative Study of Clustering Algorithms Using WEKA Tools | Analyse three major clustering algorithms and compare their performance. | k-Means clustering, Hierarchical method , Density-based clustering. | Performance of k-Means algorithm is better than the Hierarchical clustering algorithm. |
| Manish Verma, Mauly Srivastava, Neha Chack, Aful Kumar Diswar, and Nidhi Gupta | 2012 | A Comparative Study of Various Clustering Algorithms in Data Mining | Review six types of clustering techniques and these techniques are implemented and analysed using WEKA. | k-Means, Hierarchical, DB Scan, density based, OPTICS, EM. | k-Means algorithm is faster than other clustering algorithms and also produces quality clusters when using huge dataset. |
| Sunila Godara, and AmitaVerma | 2013 | Analysis of Various Clustering Algorithms | Review four types of clustering techniques with performances. | k-Means cluster, Farther First, Density based, Filtered cluster. | k-Means is faster than other clustering algorithm and also produces quality clusters. |
| Namita Bhan, and Deepthi Mehrotra | 2013 | Comparative Study of EM and k-Means Clustering Techniques in WEKA Interface | Review details about the performance of k-means and Expectation Maximization (EM) using WEKA. | k-Means and Expectation Maximization (EM). | K-Means algorithm is very bad at handling overlapping data-points while EM does much better on the overlapping data. |

| | | | | | |
|---|---|---|---|---|---|
| Pankaj Saxena, and Sushma Lehri | 2013 | Analysis of Various Clustering Algorithms of Data Mining on Health Informatics | Study the various clustering algorithms and show the comparison of different clustering algorithms of data mining and find out which algorithm suitable for users working on health data. | COBWEB<br>DBSCAN<br>Hierarchical<br>K-Means | k-Means clustering algorithm is the simplest and fastest as compared to others. |
| Garima Sehgal, and Kanw al Garg | 2014 | Comparison of Various Clustering Algorithms | Compare the algorithm according to the factors size of the dataset, numb of clusters and time taken to form clusters using WEKA. | Partitioning based:-<br>k-Means<br>Farthest First<br>EM<br>Non-partitioning:-<br>Density-based<br>Hierarchical based<br>COBWEB. | The size of datasets increase, time taken to form clusters increases.<br>In partitioning based clustering:-<br>The Farthest First took the least time in forming clusters whereas EM took maximum time.<br>In non-partitioning based:-<br>Density Based took the least time while Hierarchical took maximum time.<br>In form of clusters number:-<br>k-Means, Farthest First, Hierarchical and Density form an equal number of clusters for all three datasets. |
| Raj Bala, Sunil Sikka, and Juhi Singh | 2014 | A comparative Analysis of Clustering Algorithms | Compare and analyse four clustering algorithm in terms of efficiency and accuracy. | k-Means, Hierarchical, Expectation Maximization, Density based. | k-Means gives better results as compared to other algorithms. |
| Saman Pooja Mittal | 2014 | Comparison and Analysis of Various Clustering Method in Data Mining on Education Dataset Using the WEKA tool | Compare the performance of three major clustering algorithms on the aspect of correctly class wise cluster building ability of the algorithm using WEKA. | k-Means<br>k-Medoids<br>Hierarchical<br>Grid based | Every technique is important in his functional area and k-Means provides better results than other methods. |
| Zainal, K Sulaiman, N.F and Jali, M.Z | 2015 | An Analysis of Various Algorithms For Text Spam Classification and Clustering Using Rapid Miner and Weka | Finding the performance of various algorithms for classification and clustering SMS messages using two different tools namely Rapid Miner and Weka. | Naïve Bayes (NB) Support Vector machine (SVM), k-Nearest Neighbour (k-NN). | Weka tool gives the shortest time in executing spam classification and clustering and also gives a higher rate of accuracy which is significantly better compared to Rapid Miner. SVM is the best classifier for spam classification and k-Means is the suitable algorithms for clustering using both tools. |

From the Table 2.6, we can suggest that k-Means is the best clustering algorithm using different datasets and data document. However, each algorithm contains its own formula and functions so it will produce different results.

Sharma and Gupta (2012) proposed a hybrid algorithm for clustering Punjabi text document that uses semantic relations among words in a sentence for extracting phases. Phrases extraction creates a feature vector of the document which is used for finding similarity among all documents. There are eight steps involved (i.e. pre-processing phase, calculating term frequency of phrases, finding top k-frequent frequent phrases, finding similar documents and creating initial clusters, calculating term frequency, finding new clusters based on frequent term, document unrecognized cluster and make clusters. Experiment results showed that hybrid algorithm performs better with real time and sets.

Delany et al., (2012) investigated methods and data used in SMS spam filtering. They wanted to identify different categories of SMS spam and perform a clustering experiment on the corpus that they collected, from NUS corpus, ICT corpus, GrumbleText and WhoCallsMe website. Firstly, they did a stop-list process in the dataset to remove common functional words and then applied basic frequency-based term selection to remove terms occurring in less than three documents. Standard log based TF-IDF was used to weigh individual terms. To cluster messages, they divided the data into a flat, disjoint partition via spectral clustering method. They managed to get ten types of the cluster in spam messages.

There are published papers related to the clustering process using AIS algorithms. Nazri et al., (2009) proposed a hybrid approach that combines the Guided Agglomerative Hierarchical Clustering (GAHC) and Artificial Immune Network (AIN) for learning concept hierarchy from Malay texts called GCAIN (Guided Clustering and Artificial Immune Network). Three stages involved were pre-processing, concept hierarchical induction using GAHC and hierarchical concept learning using AIN. An ontology learning system has been built based on GAHC and GCAIN and tested on three different corpus domains. Results showed that GCAIN performs better than GAHC in all three domains and it has a greater ability to be used in learning concept hierarchies. In 2010, they again proposed a new hierarchical clustering algorithm for learning concept hierarchies named Clonal Selection Algorithm for Learning Concept Hierarchies, or CLONACH. This algorithm is a combination of an artificial immune system, named CLONACH and bisecting k-Means. An ontology learning system was built based on CLONACH and was tested on texts from three different domains. Results showed that CLONACH is better than GAHC in terms of taxonomic overlap.

Tang and Vemuri (2005) proposed the use of the aiNet (Artificial Immune Network) for document clustering. aiNet is an immune system based algorithm which combines the pre-processing and clustering procedures. The original aiNet algorithm uses Hierarchical Agglomerative Clustering (HAC) to detect clusters but then they used both HAC and k-means. Also, the Principle Component Analysis (PCA) is integrated into this method to reduce the time complexity. The result shows that aiNet is capable of obtaining better clustering results and by integrating PCA, the cluster documents perform more accuracy

than without PCA. Their approach is good for large-sized document sets that contain data redundancy and noise.

## 2.6 SUMMARY

The field of computer science not only involves in computer and programming language but also can use theory from other areas. This research applies the concept of BIS in biology to identify spam text messages. Four types of AIS algorithms (i.e. Immune Network Theory, Clonal Selection, Negative Selection and Danger Theory) are developed and since these theories have the relationship with BIS, their characteristics are suitable to be used for identifying spam text messages. Our main research focus is on the classification phase (i.e. process of clustering spam messages into categories) and two types of AIS algorithms are used - Clonal Selection and Immune Network Theory that are discussed further in the next chapter.