5th International Conference on the Advancement of Science and Technology

Computing

Proceedings

ICAST
Bahir Dar, Ethiopia

May 2017

Wisdom at the source of the Blue Nile

Bahir Dar Institute of Technology
Bahir Dar University

The 5th International Conference on the Advancement of Science and Technology

ICAST-2017

Proceedings

Faculty of Computing

Bahir Dar Institute of Technology

Bahir Dar University

May 2017

Bahir Dar, Ethiopia

EDITORS

Table of Contents

Amharic Word Sense Disambiguation Using Wordnet

Seid Hassen Yesuf[1], Yaregal Assabie[2]
[1]Department of Computer Science, Wollo University, Dessie, Ethiopia, seid2u@gmail.com
[2]Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia, yaregal.assabie@aau.edu.et

***Abstract* - Words can have more than one distinct meaning and many words can be interpreted in multiple ways depending on the context in which they occur. The process of automatically identifying the meaning of a polysemous word in a sentence is a fundamental task in Natural Language Processing. This paper presents Amharic word sense disambiguation method using Amharic WordNet. Amharic WSD extracts knowledge from word definitions and relations among words and senses that are found in Amharic WordNet. We have evaluated Amharic word sense disambiguation using WordNet system by conducting two experiments. The first one is evaluating the effect of Amharic WordNet with and without morphological analyzer and the second one is determining an optimal windows size for Amharic WSD. Amharic WordNet with morphological analyzer and Amharic WordNet without morphological analyzer we have achieved an accuracy of 57.5% and 80%, respectively. In the second experiment, we have found two-two-word window on each side of the ambiguous word is enough for Amharic WSD.**

Keywords- **Natural Language Processing, Amharic WordNet, Word Sense Disambiguation, Knowledge Based Approach**

I. INTRODUCTION

Natural Language Processing (NLP) provides tools and techniques that facilitate the implementation of natural language-based interfaces to computer systems, enabling communication in natural languages between man and machine. In all natural languages, many words can be interpreted in a variety of ways, in accordance with their context. Natural language processing (NLP) involves resolution of various types of ambiguity. Lexical ambiguity is one of these ambiguity types, and occurs when a single word (lexical form) is associated with multiple senses or meanings. Ambiguity is a major part of any human language. Almost every word in natural languages is polysemous, that is, they have numerous meanings or sentences. Word sense disambiguation involves picking the intended sense of a word for a pre-defined set of words, which is typically, a machine-readable dictionary, such as WordNet. The lexical and semantic analysis of words is necessary for computers to make sense of the words called as word sense disambiguation [1].

The word sense ambiguity is a hard problem for the developers of Natural Language Processing (NLP) systems. Words often have different meaning in various contexts. The process by which the most appropriate meaning of an occurrence of an ambiguous word is determined word sense disambiguation, and remains an open problem in NLP. Humans possess a broad and conscious understanding of the real world, and this equips them with the knowledge that is relevant to make sense disambiguation decisions effortlessly, in most cases. However, creating extensive knowledge bases, which can be used by computers to 'understand' the world and reason about word meanings, accordingly, is still an unaccomplished goal of Artificial Intelligence (AI) [3].

The WSD problem is that of associating an occurrence of an ambiguous word with one of its senses. In order to do this, first, an inventory of the senses associated with each word to be disambiguated must be available; second, a mechanism to associate word senses in context to individual senses must be developed, and thirdly, an evaluation procedure to measure how well this disambiguation mechanism performs must be adopted. Designing the actual disambiguation mechanism involves the construction of disambiguation rules and their subsequent application to a real disambiguation problem, achieving WSD. In general, people are unaware of the ambiguities in the language they use because they are very good at resolving them using their prior and textual knowledge. However, computer systems do not have this knowledge, and consequently do not do a good job of making use of the context [3].

Word Sense Disambiguation (WSD) is the task of automatically identifying the correct meaning of a

word that has multiple meanings. In WSD, these meanings are referred to as senses, or concepts, which are obtained from a sense-inventory. The ambiguous word is referred to as the target word and the context in which the target word used is called an instance. WSD is not thought of as an end in itself, however, as an enabler for other tasks and applications of computational linguistics and natural language processing (NLP) such as parsing, semantic interpretation, machine translation, information retrieval, question answering, text mining, computational advertising and the like. Amharic WSD developed by the previous researchers limited on corpus-based approach to word sense disambiguation that requires information that can be extracted from labeled and unlabeled training data. To deal with this problem knowledge-based approach to Amharic WSD technique is proposed. Knowledge-based methods use information extracted from structured data called a knowledge source. These methods rely on information from the knowledge source about a concept such as its definition or synonym rather than training instances in manually annotated or unannotated training data. Knowledge-based Amharic WSD extracts knowledge from word definitions and relations among words and senses. This work is aimed at developing Amharic word sense disambiguation using WordNet that identify ambiguous words with their contexts. The remaining part of this paper is organized as follows. Section 2 presents Amharic language with emphasis on its word ambiguity. Section 3 presents related works. While section 4 discusses Amharic word sense disambiguation using WordNet. In Section 5, we present experimental results. Conclusion and future works are presented in Section 6. References are provided at the end.

II. AMHARIC LANGUAGE AND AMHARIC WORD AMBIGUITY

A. Amharic Language

Amharic is the working language of Ethiopia, although many languages are spoken in Ethiopia. It is also the second most spoken Semitic language in the world next to Arabic [21]. Amharic is written using Ethiopic script which has 33 consonants (basic characters) out of which six other characters representing combinations of vowels and consonants that are derived for each character.

Amharic has a complex morphology. Sentences in Amharic are often short in terms of the number of words they were formed. This nature of the language makes the window size (bag of context words) narrow. In other token, context words surrounding the word have more advantage for disambiguation purpose in WSD area. Studying morphological aspects of languages helps to distinguish between lexical components of words, which are accountable for the semantics of the words, and grammatical words. Verbs are morphologically the most complex word class in Amharic with many inflectional forms. Sentences in Amharic are often short in terms of the number of words they are formed [4].

Syntactically, Amharic is an SOV language i.e. subject + object+ verb [5]. For example, the sentence in English “Abebe sharpen the knife” can be written in Amharic as አበበ ቢላዋ ሳለ ('ababa bîlãwã sãlé). We know that the Amharic word sãlé(ሳለ) has four meanings that is “sharp”, “to vow”, “to drew”, and the other is “to cough”. Therefore, it is an ambiguous word. Disambiguation can be performed to identify the sense of ambiguous word what the sentences are talking about after considering neighboring words (አበበ/'ababa/, ቢላዋ/bîlãwã/).

B. Ambiguities in Amharic Language

Ambiguity is an attribute of any concept, idea, statement or claims whose meaning, intention or interpretation cannot be definitively resolved according to a rule or process consisting of a finite number of steps. In ambiguity, specific and distinct interpretations are permitted (although some may not be immediately apparent), whereas with information that is vague, it is difficult to form any interpretation at the desired level of specificity. Researcher [6] classified Amharic word ambiguity in to six types, i.e. lexical ambiguity, phonological ambiguity, structural ambiguity, referential ambiguity, semantic ambiguity, and orthographic ambiguity.

Lexical Ambiguity: Word meanings in more than one sense can lead to different interpretations by different individuals. Lexical ambiguity comes into being when two or more of the meanings of a word are applicable in a given situation. In other words, lexical ambiguity refers to a case in which either a lexical unit belongs to different part-of-speech categories with different senses, or to a lexical unit for which there is more than one sense, while these different senses fall into the same part of-speech category [4,8]. The following examples show the synonymy, homonymy and categorical ambiguity structures of lexical ambiguity. Consider the Amharic sentences, baqlõ ’ajahú, “በቅሎ አየሁኝ” The word “baqlõ” is ambiguous since it has either a noun or a verbal meaning. When we interpret the word, it provides senses as:

1. When the word baqlõ used as noun, it means, “I saw a mule”.

2. When the word comes with verbal meaning, it means “I saw something which is grown (may be before a type of plant)”.

For instance, the word sãlé (ሳለ) has also the same spelling and sound for two meanings. Consider Amharic text “አበበ ስእል ሳለ”/ sãlé/, from this example the word sãlé (ሳለ) can be interpreted as “draw”or“to paint”. It can also be understood as “cough” based on the context words around the ambiguous word “ሳለ” in the sentence, “አበበ ጉንፋን ስለ ያዘው ሳለ”. It can also be understood as “to sharp” based on the context words around the ambiguous word “ሳለ”/sãlé / in the sentence, “አበበ ቢላዋ ሳለ”.

Semantic ambiguity: It determines the possible meanings of a sentence by focusing on the interactions among word level meanings in the sentence. It is caused by polysemy, idiomatic and metaphorical constituents of sentences [7]. The following examples show the structures of sematic ambiguity.

Phonological Ambiguity: Interpretation of speech sounds within and across words may cause ambiguity. Phonological Ambiguity occurs when the speakers pronounce by creating pause sound. Speaking using pauses and without it leads to word ambiguity [Teshome, 2008]. For example, ደግ ሰው ነበር “deg + sew neber”. In the sentence “+” sign shows where the pause is. When the sentence is pronounced with pause it senses as “He was a kind man”, however if it is pronounced without pause. It will provide different sense from the previous i.e. “They had preparation for a ceremony”.

Syntactic Ambiguity: Structural (syntactic) Ambiguity results when word order becomes in unorganized manner and holds more than one possible position in the grammatical structure of the sentence. Syntactic disambiguation of words that can function as multiple parts-of-speech is accomplished by reorganizing the order at the syntactic level [Teshome, 2008]. Consider the Amharic sentence, “የሀበሻ ታሪክ አስተማሪ” this sentence can have two different interpretations: “a person who teaches Abyssinian history” and “an Abyssinian who teaches history”. It can be further illustrated using structural organization of the sub-constituent/tarik/ ‘History’.

Referential Ambiguity: This ambiguity arises when a pronoun stands for more than one possible antecedent. For example, “ካሳ ስለተመረቀ ተደሰተ” Kasa sletemereke tedesete. In Amharic, pronoun is understood by default even if it is not written grammatically. This sentence has two different readings. I.e. Kasa was pleased himself because he graduated. “ካሳ ስለተመረቀ ራሱ ተደሰተ።” and Somebody was pleased because Kasa graduated. “ካሳ ስለተመረቀ ተደሰተ።”

Orthographic Ambiguity: Orthographic Ambiguity results from geminate and non-geminate sounds. For example, liju yislal “ልጁ ይስላል’’. The word “yislal” is the cause of ambiguity. The sentence is ambiguous between the following meanings. He draws (“yslal”) and He coughs (“yslal “).

III. RELATED WORK

Corpus-based approaches attempt to disambiguate words by using information gained from training on some corpus, rather than taking it directly from an explicit knowledge source. Training can be carried out either on a disambiguated corpus or a raw corpus. In a disambiguated corpus, the semantics of each polysemous lexical item has been marked, while in a raw corpus, the semantics has not been marked yet

[20]. Knowledge-based disambiguation is carried out by using information from an explicit lexicon or knowledge base. The lexicon may be a machine readable dictionary, thesaurus or WordNet used as the knowledge-base to disambiguate word senses

The author [1] has done on WSD for the Amharic language. The researcher has studied how linguistic disambiguation can improve the effectiveness of an Amharic document query retrieval algorithm. The author developed Amharic disambiguation algorithm based on the principles of semantic vectors analysis for improving the precision and recall measurements of information retrieval for Amharic legal texts and implemented in Java. The researcher used the Ethiopian Penal Code which is composed of 865 Articles was used as a corpus in the study. The disambiguation algorithm was used to develop a document search engine. The researcher developed an algorithm based on distributional hypothesis stating that words with similar meanings tend to occur in similar contexts. For disambiguation of a given word, the author computed the context vector of each occurrence of the words. The context vector was derived from the sum of the thesaurus vectors of the context words. The author constructed the thesaurus by associating each word with its nearest neighbors.

However, for evaluating WSD the author used pseudo words, which are artificial words, rather than real sense tagged words reasoning that it is costly to prepare sense annotation data. The researcher compared the developed algorithm with Lucene algorithm and reported that the algorithm is superior over the Lucene's one. The researcher achieved 58%-82% average precision and recall, respectively.

The author [9] used corpus based, supervised machine-learning approach using Naïve Bayes algorithm for Amharic WSD, which is used to check standard optimal context window size, which refers to the number of surrounding words sufficient for extracting useful disambiguation. Based on Naïve Bayes algorithms, experiment found that three-word window on each side of the ambiguous word is enough for disambiguation. The author used a monolingual corpus of English language to acquire sense examples and the sense examples are translated back to Amharic, which is one approach of tackling the knowledge acquisition bottleneck. Based on Naïve Bayes algorithm, the experiments were conducted on WEKA package. The author concluded that, Naïve Bayes methods achieve higher accuracy on the task of WSD for selected ambiguous words and achieved accuracy within the range of 70% to 83% for all classifiers. This is an impressive accuracy for supervised WSD; however, it suffers from knowledge acquisition bottleneck. However, supervised machine learning approach of WSD performs better by human intervention; however, this research has limitations of knowledge-acquisition bottleneck, i.e., it requires manually labeled sense examples which takes much time, very laborious and therefore very expensive to create when the corpus size increases.

The author [10] used a corpus-based approach to word sense disambiguation that only requires information that can be extracted automatically from untagged text. Unsupervised machine learning technique was applied to address the problem of automatically deciding the correct sense of an ambiguous word. The author used corpus of Amharic sentences, based on five selected ambiguous words to acquire disambiguation information automatically. A total of 1045 English sense examples for five ambiguous words were collected from British National Corpus (BNC). The sense examples were translated to Amharic using Amharic-English dictionary. The author tested five clustering algorithms: simple k-means, hierarchical agglomerative: single, average and complete link and expectation maximization algorithms, in the WEKA package. Based on the selected algorithms, the author concluded that simple k-means and EM clustering algorithms achieved higher accuracy on the task of WSD for selected ambiguous words. The author achieved accuracy within the range of 65.1 to 79.4 % for simple k means, 67.9 to 76.9 for EM and 54.4 to 71.1 for complete link clustering algorithms for five ambiguous words. The strength of this research is that the context of an ambiguous word is clustered in to a number of groups and discriminate these groups without actually labeling them. However, the limitation of this research is that training data is required for each word that need to be

disambiguated and unsupervised method cannot rely on a shared reference inventory of sense. The approach used by the researcher is that the same sense of a particular word will have neighbor words alike and clustering word occurrences and classifying new occurrences into induced clusters. The author [11] has implemented the other WSD for Amharic language. The author used Semi supervised method for five words only. To disambiguate words, the author used abundant unlabeled training data and limited labeled training data, because labeling the training examples requires human efforts that are costly. So, semi-supervised learning which tries to exploit many unlabeled examples with some seed examples to improve learning performance. Although seed examples selection was a challenging task, semi-supervised learning that tries to exploit many unlabeled examples with some seed examples to improve learning has been implemented for senses disambiguation task in semi- supervised study. The author used WEKA package for clustering and classifying data. The problem of word ambiguity in Amharic is being tried to be solved by preparing a five selected words corpus after a total of 1031Amharic sentences were collected. Two clustering algorithms, expected maximization and k-mean, were employed for clustering of sentences into their senses. The average performance of the employed on five classifying algorithms specifically AdaboostM1, bagging, ADtree, SMO, and Naïve Bayes were 83.94%, 78.28%, 88.47%, 87.40% and 47.98% respectively. The strength of this research is that unlike supervised approach, semi-supervised approach needs only a few seeds instead of a large number of training examples. However, the limitation of this research is that training data is required for each word that needs to be disambiguated.

On the other hand, Tesfa [12] used corpus based approach to disambiguation where supervised machine learning techniques were applied to a corpus of Afaan Oromo language, to acquire disambiguation information automatically. The researcher also applied Naïve Baye's method to find the prior probability and likelihood ratio of the sense in the given context. The Author used corpus of Afaan Oromo sentence based on five selected ambiguous words to acquire disambiguation information automatically a total of 1240 Afaan Oromo sense examples were collected for selected five ambiguous words and the sense examples were manually tagged with their correct senses. A standard approach to WSD is to consider the context of the ambiguous word and use the information from its neighboring or collocation words. The contextual features used in this study were co-occurrence feature, which indicate word occurrence within some number of words to the left or right of the ambiguous word and k-fold cross-validation statistical technique was applied for performance evaluation. However, supervised machine learning approach of WSD performs better by human intervention; however, this research has limitations of knowledge-acquisition bottleneck, i.e., it requires manually labeled sense examples which takes much time, very laborious and very expensive to create when the corpus size increases and training data is required. The researcher achieved an accuracy of 79% and found four-word window on each side of the ambiguous word is enough for Afaan Oromo WSD.

Seo et al. [13] also did unsupervised word sense disambiguation using WordNet for English language. Word sense disambiguation method for a polysemous target noun using the context words surrounding the target noun and its WordNet relative's words, such as synonyms, hypernyms and hyponyms used to disambiguate. The result of sense disambiguation is a relative that can substitute for that target noun in a context. The selection was made based on co-occurrence frequency between candidate relatives and each word in the context. Since the co-occurrence frequency is obtainable from a raw corpus, the researchers used unsupervised learning algorithm, unsupervised learning use a raw corpus and therefore does not require a sense-tagged corpus. Finally, the researchers evaluated the developed system on 186 documents in Brown Corpus and achieved 52.34%of recall and the researchers do not consider a way to utilize the similarity between definitions of words in WordNet.

Other researchers [14] have also implemented word sense disambiguation for the English language

using a knowledge-based approach. The authors proposed a robust knowledge-based solution to the word sense disambiguation problem for English language. It is observed that ambiguous words are resolved using the word's part-of-speech and the contextual information found in the sentence. The researchers resolving an ambiguous word based on the word's POS is possible when the parse tree is unambiguous. However, problems may arise when multiple parse trees can be formed due to the absence of an optional term and the presence of a term with an ambiguous POS.

These previous studies have limitation of less data, less coverage of ambiguous words and ambiguity types and corpus was used as a source of information for disambiguation. In our study, Amharic WordNet is used as a source of information for disambiguation. The system disambiguates words in running text, referred to as all-words disambiguation due to lexical-sample methods can only disambiguate words in which there exists on a sample set of training data in which ambiguous words may not be known ahead of time and the real sense of ambiguous word is retrieved from Amharic WordNet.

IV. ARCHITECTURE OF AMHARIC WORD SENSE DISAMBIGUATION USING WORDNET

In our study, Amharic WordNet is used as a source of information for disambiguation and knowledge-based Amharic WSD method that allows the system to disambiguate words in running text, which is called all-words disambiguation. All-words disambiguation methods have an advantage over what is termed lexical-sample disambiguation methods that was done by the previous researchers. Lexical sample methods can only disambiguate words in which there exists a set of training data where ambiguous words may not be known ahead of time. We determine the correct concept of ambiguous words by first identifying the ambiguous words semantic type, which is a broad categorization of a concept. After the semantic type of the ambiguous words is identified, then the correct concept is identified based on its semantic type from Amharic WordNet.

The system architecture of the proposed Amharic word sense disambiguation system is composed of four essential components namely, preprocessing component, morphological analysis component, Amharic WordNet database and disambiguation component. Fig. 1 illustrates the Architecture of Amharic word sense disambiguation using WordNet. The system takes text as an input and identifies the ambiguous words and its sense form Amharic WordNet. The texts are preprocessed to make suitable for further processing. Morphological analysis is important for morphologically complex languages like Amharic because it is impossible to store all possible words in WordNet. We used morphological analysis to reduce various forms of a word to a single root word. Morphological analysis produces root word and provides the root word-to-word sense disambiguation component particularly to the ambiguous word identification. The Amharic WordNet database contains Amharic words along with their different meanings, synsets and semantic relations within concepts. This component helps to implement the components of WSD. Word sense disambiguation component is responsible to identify the ambiguous word and to assign the appropriate sense to ambiguous word. To accomplish this, it incorporates various components such as Ambiguous Word Identification, Context Selection, Sense Selection and Sense retrieval components. The WSD components are integrated with Amharic WordNet.

Preprocessing tasks are data preparation procedures that should be done before dealing with different text mining techniques. Pre-processing is involved in preparing the input sentence into a format that is suitable for the morphological analysis. The pre-processing stage consists of steps such as tokenization, normalization and stop-word removal. The tokenization takes the input text supplied from a user and tokenizes it into a sequence of tokens, which is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called token and finally it gives the tokens to the next phase. For most languages, whitespaces and punctuation

marks are used as boundary markers. The Amharic language has its own punctuation marks which demarcate words in a stream of characters which includes ‘hulet neTb’ (:), ‘arat neTb ‘(: :), ‘derib sereze’ (፤), ‘netela serez’ (፣), exclamation mark ‘!’ and question mark’?’ These punctuation marks do not have any relevance to identify the meaning of ambiguous words using WSD. Therefore, except ‘arat neTb ‘and ‘question mark’ which are used to detect the end of the sentence, all other punctuations is detached from words in tokenization process. Normalization is performed on the word tokens that result from text segmentation. In the Amharic language, two types of normalization issues arise [46]. The first one is the identification and replacement of shorter forms of a word that is written using forward slash “/” or period “.”. An example is the replacement of “መ/ር” by “መምህር”. The second normalization issue is the identification and replacement of Amharic alphabets that have the same use and pronunciation, but they have different representations of alphabets. The replacement is made using a representative alphabet from a set of similar alphabets. For example, the word “cough” can have two representations in Amharic: “ሳለ” and “ሣለ”. These two words differ only by their first characters: “ሳ” and “ሣ” and have similar usages and different forms. They need to be converted to a single representative character such as “ሳ” [15]. Stop words are low information bearing words such as “ነው”or “ና”, typically appearing with high frequency. In our approach, “stop words” like ‘ነው’, ‘እስከ’, ‘እንደ’, etc. are discarded from input texts as these words are meaningless to derive the “sense” of the particular sentence. Then, the text containing meaningful words (excluding the stop words) pass through morphological analysis.

Morphological analysis is the segmentation of words into their component morphemes and the assignment of grammatical morphemes to grammatical categories and lexical morphemes to lexemes. Morphological analysis is important for morphologically complex languages like Amharic because it is practically impossible to store all possible words in a lexicon. This becomes obvious in the context of machine translation to a morphologically simple language such as English, where the correspondence between words in Amharic, Oromo, or Tigrinya and the other language will often be many-to-one [16]. Amharic root words can generate hundreds of lexical forms of different meanings. Morphological analysis is used for finding the root morphemes of the words and it is a one of the components used to reduce various original forms of a word to a single root or stem of these words. It is necessary to represent different word forms in a single format and to reduce memory usage for storing the words. In morphologically complex language like Amharic, a morphological analysis will lead to significant improvements in WSD systems.

WordNet for Amharic language has a significant impact on search engine, automatic text categorization and Amharic word sense disambiguation [17]. In Amharic WordNet, the words are grouped together according to their similarity of meanings. Amharic WordNet is a system for bringing together different lexical and semantic relations between the Amharic words. Amharic WordNet (AWN) is used to identify the ambiguous word and contains a list of senses for given words from the input sentence.

Amharic word sense disambiguation using composed of context selection, ambiguous word identification, sense selection and sense retrieval components. Ambiguous word identification is a component used to identify the ambiguous word from the input sentence based on information provided on Amharic WordNet. The ambiguous word identification is to be checked whether each root word exists in the Amharic WordNet or not. Words that are found in Amharic WordNet have their own sense on AWN. If words do not exist in AWN, the word is discarded. For example, if the following sentence is the input sentence: “ተማሪው ጉንፋን ስለያዘው ሳለ።” First, the input sentence is preprocessed. After morphological analysis, only four words will be left (i.e.ምእር, ጉንፋን, ይእዝ and ስእል) in the input sentence. Then each root word with respect to its sense in the input sentence is counted in AWN. In our case, the root word “ስ-እ-ል” and its sense exist five times in WordNet. So that, “ስ-እ-ል” is detected as ambiguous word in the input

sentence and “ስ-እ-ል” is the root word for the word “ሳለ”. Therefore, “ሳለ” is ambiguous word and their context of the sentence. The context in WSD refers to the words surrounding the ambiguous words, which are used to decide the meaning of the ambiguous word. For example, the sentence “ተማሪው ጉንፋን ስለያዘው ሳለ።” after morphological analysis, it will be “ም-እ-ር, ጉንፋን, ይ-እ-ዝ and ስ-እ-ል”. From the sentence the ambiguous word is “ሳለ” based on ambiguous word identification component and the contexts are the surrounding words or neighbor words on the ambiguous word “ሳለ” which are {ም-እ-ር, ጉንፋን, ይ-እ-ዝ} its senses are identified using the ontology based related words and a sense overlap. Therefore, the correct sense of a word is obtained from the context of the sentence. Context selection uses the words of the sentence itself as Sense selection chooses senses that lead to highest overlap and related words and sense is retrieved from Amharic WordNet using sense retrieval component. For example, “አሾለ፥ሞቁረጫ ጠርዝን አተባ” is the sense of ambiguous word for “ሳለ”, for the input sentence “አበበ ቢላዋ ሳለ።” to identify whether the given sense is as a sense of the given ambiguous word. First, the context of ambiguous word is identified by using context selection

.

sense is retrieved in Amharic WordNet based on the

context, including ambiguous words and selects the context that contains ambiguous words from Amharic WordNet. The sense selection component finds the number of overlapping of the words from the set of words output by the context selection component with the sense of ambiguous word. The root word definition with the ambiguous words having highest overlapping is selected as the sense of the ambiguous word. For each of the root words in sense selection set, all sense of root word and ambiguous words were identified in the entire Amharic WordNet and identify all sense definitions of the words to be disambiguated from AWN and determine the definition overlap for all possible sense combinations.

component and sense overlap. Second, the context of the ambiguous word based on the input sentence is selected in this case “ቢላዋ” the context of the ambiguous word. Finally, the sense overlap of “ቢላዋ” and the ambiguous word “ሳለ” are selected in Amharic WordNet. The sense retrieval component is responsible for extracting a sense of ambiguous word from Amharic WordNet, and associating the extracted sense with the ambiguous words

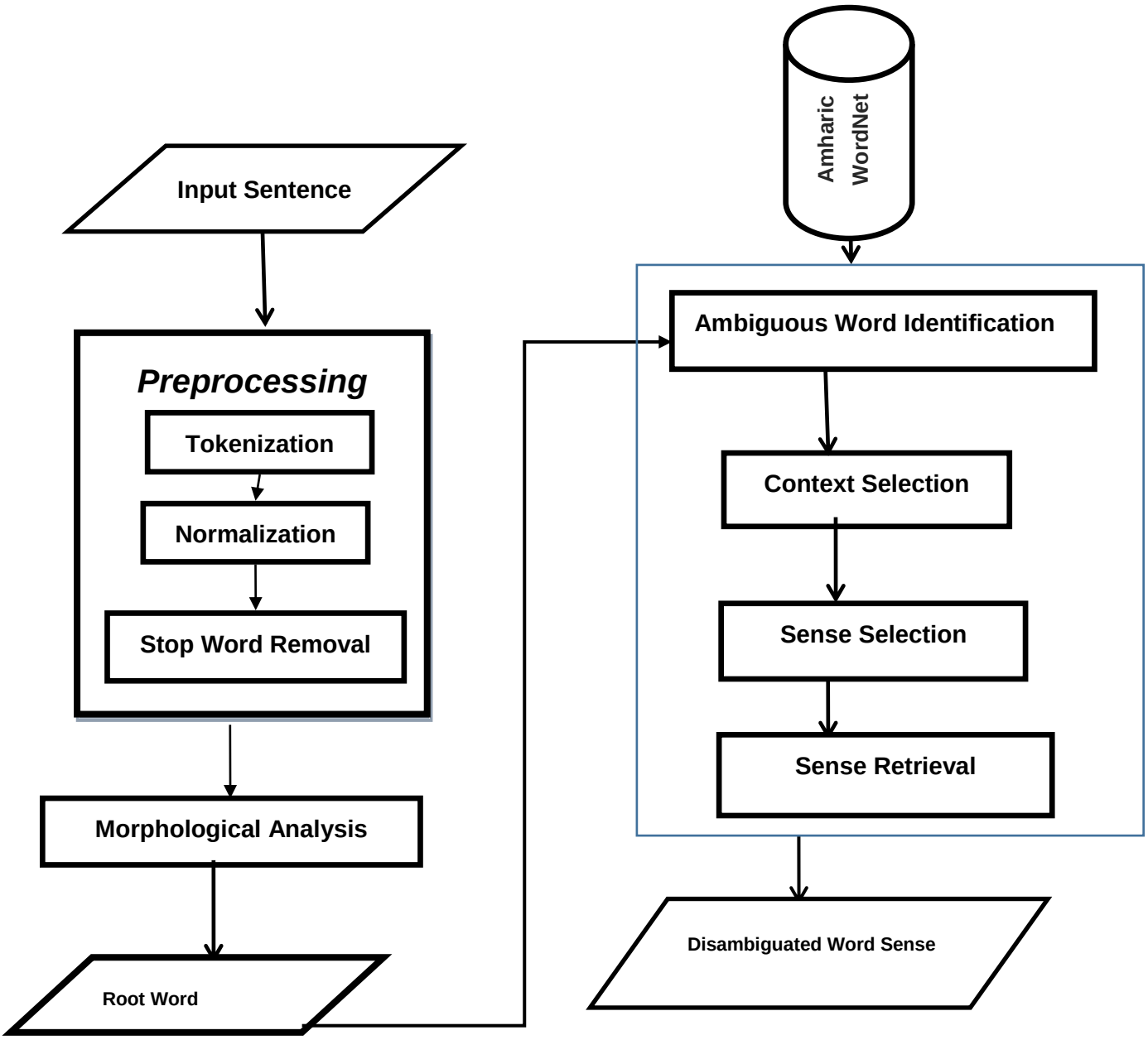


FIGURE 1: ARCHITECTURE OF AMHARIC WORD SENSE DISAMBIGUATION USING WORDNET

V. EXPERIMENT

A. Amharic WordNet

Knowledge-based Amharic WSD method that does not require sense tagged corpus and that identifies senses of all words in sentences or not a small number of words. Our proposed method depends on Amharic WordNet, which is relatively very large, and it is a lexical database in a hierarchical structure. The development of Amharic WordNet is an important step for WSD, even for other application areas such as Information Retrieval, Machine Translation and soon.

The major source of the data we used for developing Amharic WordNet was from Amharic dictionary [18] and words selected by a linguistic expert from a list of homonyms collected by Girma [19]. The Amharic WordNet contains 10,000 synsets and 2000 words. We performed an evaluation of the proposed Amharic WSD algorithm using the context window and morphological analyzer effect for word sense disambiguation. A test sentence of 200 random sentences containing the ambiguous words from the knowledge base is created. Some sentences are taken from linguistic experts and some are taken from newspapers. Out of several senses for each ambiguous word, we considered only two or three senses that are most frequently used. To evaluate the performance of Amharic word sense disambiguation we used precision and recall.

B. Test Results

In this study, two experiments were conducted. The first experiment was conducted on Amharic WordNet with and without morphological analyzer since knowledge-based methods do not use any manually or automatically generated training data, however, use information from an external knowledge source so that the sense inventory for these methods comes from the knowledge source being used were Amharic WordNet. This experiment is performed to test whether morphological analyzer improve the performance of Amharic WSD or not. We have carried out a number of evaluations of WSD algorithms using different linguistic resources in various combinations. A total number of 200 sentences are used to evaluate performance of the system. The second experiment is investigating the effect of different context sizes on disambiguation accuracy for Amharic to point out the optimal window size. For the purpose, different dimensionally variant data sets were tested starting from 1-left and 1-right to 5-left and5-rightwindow sizes.

Table 1: Performance of the Amharic WSD using Amharic WordNet system with and without Morphological Analyzer

	With morphological analyzer	Without morphological analyzer
Recall	90.3%	74.5%
Precision	84.8%	66.67%,
F-Measures	87.5%	70.37%
Accuracy	80%	57.5%,

As shown in Table 1, we can say that for all word sense disambiguation tasks, Amharic WordNet with morphological analyzer improved the accuracy of knowledge based word sense disambiguation. As we can see Amharic WordNet with morphological analyzer the context of word is found by determine the meaning of a word by using ontology based related words and the overlap of the sense of target word to each word, we achieved an accuracy of 80%. Morphological analyzer reduces various forms of word into their common root or stem word. This minimizes the consideration of the variants of a word as different word by WSD. If morphological analyzer is done, the variant of a word is taken as the same pattern, which will improve the accuracy of the knowledge based word sense disambiguation algorithms.

Table 2: Summary of experiment in different window sizes

Window size	1 -1	2 -2	3 -3	4 -4	-5
Precision	78.5%	77.6%	66.5 %	64.7%	5.7 %
Recall	70.5%	84.7%	87.4%	89.8%	0.4 %
F1-Measure	74.28%	81%	75.5%	75.2%	8.4 %
Accuracy	71.5 %	86.5 %	80.4 %	78.2%	9.9 %

An optimal context window size refers to the number of surrounding contexts for sense disambiguation, which is obtained through researches. For example, in English, a standard two-two-word window (2-2) on either side of the ambiguous word is found to be enough for disambiguation [20]. Solomon Mekonnen [9] reported that Window size of three-three words (3-3) is considered to be effective using supervised learning method with achieved accuracy within the range of 70 to 83% on five ambiguous words (መሰሰት, መጥራት, ቀረፀ,አጠና and መሰል). Solomon Assemu [10] tested the optimal window size using unsupervised learning methods and the author advised that window size of 3-3 or 2-2 is enough for disambiguation depending on the algorithms used. Window size of 3-3 was effective for Simple K means and EM clustering algorithms achieved accuracy ranged from 65.1 to 76.9 % whereas windows 2-2 was effective for agglomerative SL and CL clustering algorithms achieved accuracy range from51.9 to 71.1% on the same five ambiguous words (መሰሰት, መጥራት, ቀረፀ,አጠና andመሰል). Getahun Wassie [11] tested the optimal window size using semi-supervised learning methods and the author advised that window size of 3-3 or 2-2 is enough for disambiguation depending on the algorithms used. Window size of 3-3 was effective for bootstrapping algorithms (adabostM1, ADtree, and bagging) with achieved accuracy of 84.90%, 81.25% and 88.45% respectively while windows 2-2 was effective for Naïve Bayes and SMO algorithms achieved an accuracyof67.01% and 87.89% respectively on five ambiguous words (አጠና, ደረሰ, ተነሳ, አለ and በላ). For this study, an experiment is carried out to test 1-1 window up to 5-5 window on both side of the target word for some ambiguous words. The previous, researches does no evaluate the precision and recall of the window size and has not been tested for Amharic words using Knowledge based approaches.

As shown in the Table 2, for the knowledge based Amharic word sense disambiguation the maximum accuracy of, precision and recall achieved 86.5 %, 84.7% and 77.6% on two-two-word window size respectively. So that to the determining the optimal window size for all word sense disambiguation is that, a small window size tends to result in high precision, however low recall. In terms of recall, if there are more words in the context, the chance of finding related word with ambiguous word at least one of them is higher and hence increased window size would lead to a higher recall. We observed especially good results for window size 2. This is because for window size=2, we can assign the sense to the first instance in a sentence. For example, in the sentence “አበበ ቢላዋ ሳለ”, if window size=2 and the target is“ሳለ”, since there is no word in the right context, we assigns the sense to the target word. This induces enormous sense of resulting in good precision for window size = 2. From the above sentence, “ቢላዋ” is in the context while disambiguating the target word “ሳለ”. We define a window size around the target polysemous word and calculate the number of words in that window that overlap with each sense of the target polysemous word. The Performance of knowledge Based methods is high and also they do not face the challenge of new knowledge acquisition since there is no training data required. To compare the test results with the previous researches [1, 9, 10] used stemmer to improve the performance of WSD, however when we compare to morphological analyzer with the stemmer. The stemmer does not improve the performance of WSD than morphological analyzer.

VI. CONCLUSION AND FUTURE WORKS

Amharic is morphologically complex and less-resourced language. This complexity poses difficulty in the development of Amharic natural language processing applications. Amharic word sense disambiguation also suffers from this problem. This research work is the first attempt to develop a word sense disambiguation system for Amharic language using Amharic WordNet. Since there are no linguistic resources prepared i.e. WordNet, Thesaurus, Machine Readable Dictionaries and others for Amharic Language, which is important for WSD purpose, we prepared Amharic WordNet manually for this study. During this preparation, we have selected 2000 words including ambiguous words. Based on these

ambiguous words, we extracted Amharic sentences from newspaper as test set by the help of language experts. The architecture of the system includes preprocessing, morphological analysis, and Amharic WordNet database and word sense disambiguation components. We can conclude that Amharic WordNet with morphological analyzer can enhance the accuracy of Amharic word sense disambiguation and two-word window on each side of the ambiguous word is enough for Amharic WSD.

Thus future works that we recommended the development of Amharic thesaurus, Amharic machine readable dictionaries and Amharic ontology used to enhance Amharic word sense disambiguation. In addition to knowledge based and corpus-based approach, there is also hybrid approach need to be investigated for Amharic WSD systems. Finally, we have used manually developed Amharic WordNet, word-synsets pairs and relationships. No full-fledged Amharic WordNet is available and constructing it manually is tedious. Developing Amharic WordNet for WSD and other Amharic natural language processing applications is important and time efficient.

REFERENCES

- Teshome Kassie (2008). Word Sense Disambiguation for Amharic Text Retrieval: A Case Study for Legal Documents, Unpublished Master's Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
- Shruti Ranjan Satapathy (2013). Word Sense Disambiguation, Unpublished Master's Thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India.
- Rada Mihalcea, E. Agirre and P. Edmonds Eds (2007). Word Sense Disambiguation Algorithms and Applications Text, Speech and Language Technology, Springer, VOLUME 33, Université de Provence and CNRS, France.
- Solomon Teferra Abate and Wolfgang Menzel (2005). Syllable-Based Speech Recognition for Amharic, University of Hamburg, Department of Informatik. Vogt-Kölln-Strasse, 30, D-22527 Hamburg, Germany.
- Daniel Gochel Agonafer (2003). An Integrated Approach to Automatic Complex Sentence Parsing for Amharic Text, Unpublished Master's Thesis, Department Of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.
- Getahun A. (2001). Towards the Analysis of Ambiguity in Amharic, JES Vol XXXIV No 2.
- Saba Amsalu Teserra (2007). Bilingual Word and Chunk Alignment: A Hybrid System for Amharic and English, Unpublished Master's Thesis, Universitat Bielefeld, UK.
- Dawit Bekele (2003), The Development and Dissemination of Ethiopic Standards and Software Localization for Ethiopia, The ICT Capacity Building Programme of the Capacity Building Ministry of the FDRE and United Nations Economic Commission for Africa, Addis Ababa, Ethiopia.
- Solomon Mekonen (2010). Word Sense Disambiguation for Amharic Text: A Machine Learning Approach, Unpublished Master's Thesis, Department of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.
- Solomon Assemu (2011). Unsupervised Machine Learning Approach for Word Sense Disambiguation to Amharic Words, Unpublished Master's Thesis, Department of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.
- Getahun Wassie (2012). Semi-supervised Machine Learning Approach for Word Sense Disambiguation to Amharic Words, Unpublished Master's Thesis, Department of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.
- Tesfa Kebede (2013). Word Sense Disambiguation for Afaan Oromo Language, Unpublished Master's Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.
- Hee-Cheol Seo, Hoojung Chung, Hae-Chang Rim, SungHyon MyaengandSoo-Hong Kim (2004), unsupervised word sense disambiguation using WordNet relatives, Department of Computer

- Science and Engineering, Korea University, Published.
- W. Faris and K.H. Cheng (2013). A Knowledge-Based Approach to Word Sense Disambiguation Computer Science Department, University of Houston, Houston, Texas, USA.
- Tessema Mindaye (2007). Design and Implementation ofAmharic Search Engine, Unpublished Master's Thesis, Addis AbabaUniversity, Addis Ababa, Ethiopia.
- Gasser M. (2012). HornMorpho: A System for morphological processing of Amharic, Oromo, and Tigrinya.Conference on Human Language Technology for Development, Alexandria, Egypt.
- Tessema Mindaye, Meron Sahlemariam and Teshome Kassie (2010). The Need for Amharic WordNet,the 5th International Conference of the Global WordNet Association,Mumbai, India.
- የኢትዮጵያ ቋንቋዎች ጥናትና ምርምር ማዕከል (1993)። አማርኛ መዝገበ ቃላት፣አዲስ አበባ፣ አርቲስቲክ ማተሚያ ቤት፣ኢትዮጵያ።
- Girma Getahun (2007), በአማርኛ ሥርዓተ-ጽሕፈት ውስጥ ድምፀ-ምክሸ ሆሄያት አጠቃቀም ማስታወሻ. Retrieved on 15 November, 2013 from: http://www.nlp.amharic.org/resources/lexical/word-lists/homonyms

Recognition of Amharic Braille Documents

Ebrahim Chekol Jibril[1] and Million Meshesha[2]
[1] Department of Computer Engineering, Istanbul Technical University, Turkey
Informatics Faculty, Addis Ababa University, Ethiopia
jibril@itu.edu.tr, [12] million.meshesha@aau.edu.et [2]

Abstract*-This paper presents a system for a design and implementation of optical Braille recognition (OBR) system for real life single sided Amharic Braille documents. The system is implemented using artificial neural network for 238 Amharic characters, 10 Arabic numerals and 19 punctuation marks. Gaussian filtering with adaptive histogram equalization and morphological operation are used to detect and remove the noises in the real life Amharic Braille documents. The noise detection and removal, and thresholding algorithms are integrated with the previous algorithm implemented for recognition of clean Amharic Braille documents. The implanted algorithms achieved an accuracy of 95.5%, 95.5%, 90.5%, and 65% for *clean, small level noisy, medium level noisy and high level noisy Braille documents respectively.

Key words **–Braille, Gaussian, OBR, Neural Network**

I. INTRODUCTION

There are more than half million visually impaired people in Ethiopia consisting of students, teachers, artists, lawyers, and also employees who have significant contribution in politics, religious, economics and social affairs of the society. For these people Braille is the means for codifying their knowledge [2]. Since 1924, there are a significant number of old Braille documents produced and used by visually impaired society throughout the country. From this, very insignificant number of Braille has actually reached to the vision society creating communication gap between the vision and visually impaired people.

Each Latin Braille character or ”cell” is made up of six dot positions, arranged in a rectangle containing two columns of three dots each, 63 possible characters and a space (none dots) are available by using any one or a combination of dots [6] [11]. Braille dots are numbered 1 through 3 on the left side of the cell, and 4 through 6 on the right side, as shown in figure 1. A typical Braille page is 280X292 mm with 40 characters long and 25 lines [1]. The horizontal and vertical distance between dots in a character, the distance between cells representing a word and the distance between two lines are specified by the Library of Congress [1].

The Braille has been adopted to Amharic, English, Arabic, and other different languages in the world, and also used for mathematical and musical notation. There are many factors affecting an OCR system. Some of them include: the mode of writing (handwritten or using Braille printer), the types of Braille paper (gray or light yellow; single side or double side), the presence of artifacts like extraneous markings, and the resolution and lighting during scanning. Optical Braille Recognition (OBR) system allows visual people to read volumes of handwritten or using printer

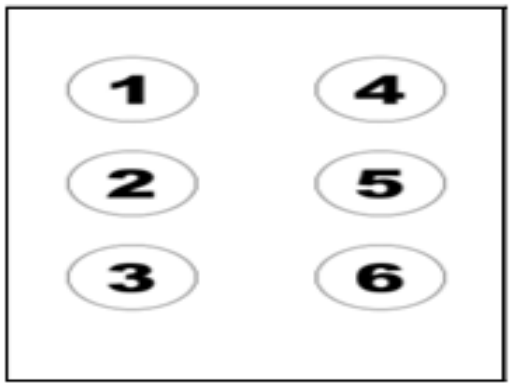


Figure 1. Dot position in a Braille Cell

Braille documents with the help of flatbed scanners and OBR software. This paper aims at developing algorithms to recognize a real life Amharic Braille documents.

II. LITERATURE REVIEW

Braille is a system of embossed (raised) signs, which are formed by six (or eight) dots. Six dots Braille is the widely used one and hence it is also considered in this study [8].

All the Amharic Braille characters, except the 6th forms, and Amharic numerals require more than one Braille cells [15][16]. Braille documents can have the dots embossed single side or double side. As the volume of Braille documents is quite large, it is advantageous to use both sides of the sheet. However, this is not the practice with Amharic Braille [7] [15].

Image filtering is a preprocessing before most image processing operations such as encoding, recognition, compression, tracking, edge detection and noise reduction [30]. Gaussian and adaptive median filtering are a well-known technique for signal smoothing [5][7]. The degree of smoothing

is determined by the standard deviation of the Gaussian [14].

Ritchings [12] work on double sided Latin characters and Arabic numbers. The documents are scanned using flatbed scanner. They achieve an average of just over 98.5% of the protrusions and 97.6% of the depressions. The majority of the errors can be attributed to the quality of the image of the Braille document. Very old documents with some of the protrusions flattened due to heavy use will give rise to more incorrectly recognized characters.

C. Ng, et.al. [4] tried to recognize both single and double sided documents written in Latin characters and Arabic numbers. The image is captured using a digital camera placed above the Braille page. They used Gaussian filtering and edge enhancement as a preprocessing technique. The filtered Braille images are binarized using a gray level histogram of the page to select a threshold value and segmented to identify characters in the Braille image. Then search the bit string against the Braille dictionary.

They used English dictionary as a post-processing. The system recognizes 100% for single sided and 97% double sided Braille images.

C. Ng, et.al, [3] attempted to develop both single and double sided recognition system for English/Chinese Braille documents. They acquire the image using digital camera. From the binarized images the researchers are applied a Gaussian noise filtering and Sobel kernels to sharpen the fine details of the image that are blurred. The bit strings of the cells are searched against the Braille dictionary, and the retrieved characters are grouped into words. Each word is then check against an English dictionary. The system performs an accuracy of 100% and 97% for single sided and double sided Braille documents respectively.

AbdulMalik et.al. [1], attempts to design Arabic optical Braille recognition system. This work tries to recognize both single side and double side documents with flatbed scanner. Cropping the image frame and de-skewed as a preprocessing technique. The proposed system is tested using variation of Braille documents: skewed, reversed, or worn-out of both single and double sided documents and they found an overall accuracy of 99%.

Néstor Falcón et.al.[10] presented the development of BrailLector, a system able to speak from Braille writing. By means of dynamic thresholding, adaptive Braille grid, recovery dots techniques and TTS software (Text-To-Speech). It is applied for single and double sided Braille images. The Braille image is acquired using a flat-bed scanner. De-skewing is used as a preprocessing technique. This final output can be presented in different formats such as a text file, a new Braille printed copy, voice or even mp3 audio format. The global image processing algorithms (dynamic thresholding, pattern detection, Braille grid creation and dots recovery using Braille grid) is very fast and robust and perform an accuracy of 99.9% for double sided Braille documents.

Teshome [13] attempt to develop Amharic Braille recognition system that enables to recognize optically scanned single-sided. The Braille documents are digitized using a flat-bed scanner. The researchers applied a global thresholding technique for image thresholding. Segmenting the Braille image is performed in two steps: first, gridlines mesh is constructed; and then dots are searched following the grids mesh with threshold. Second, clearly display the dot with single white pixel on black background. For feature extraction they adopted modification of region based approach. Feed forward neural network is applied on the extracted features to recognize the characters. The proposed system is tested on single sided and clean Braille documents and found an accuracy of 92.5%.

Currently, Braille recognition research is at a high level; most of the researches are not used image noise detection and removal techniques. However, in the practical image acquisition systems and conditions, noise is a common phenomenon in the acquisition process and also Braille noise is common due to repeated use, bend and scratch. As a result, these artifacts significantly affect the subsequent recognition process. So to address the problem noise detection and removal techniques need to be applied on the resulting image. Recognition of Braille images that are corrupted by noises has been the goal of the present research.

III. METHODOLOGY/THE SYSTEM

The preprocessing task has been done using MATLAB. Microsoft Visual C++ programming language, which is used to integrate the preprocessed image to segment and extract the features of Braille image. Artificial Neural Network is used for the purpose of classification, which is categorized as one of the robust approaches to deal with uncertainty and noisy input data [13]. For the classification purpose, MATLAB Neural Network toolbox is used mainly due to its availability and ease of constructing and training the network.

The process of Braille document recognition passes through many stages starting from Braille acquisition to recognition of Amharic characters (Figure 2).

1. Image Acquisition

We have used flat-bed scanner at 200 dpi resolution for image acquisition because it is a cheap alternative and it is easy and quick to use [10]. The 267 characters from the fourth version Amharic Braille are used for training and from the collected document, 200 characters for each image type (i.e. clean, small noise, medium and high noise) are digitized for testing purpose. The digitized image is saved in both gray and colored (RGB) window bitmap format.

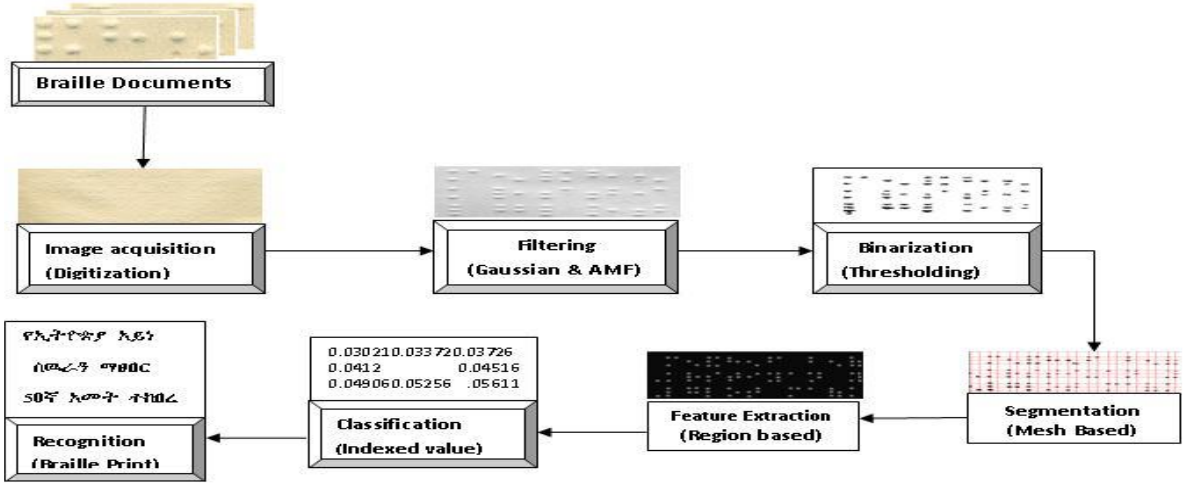


Figure 2. Block Diagram of the Amharic Braille Recognition system

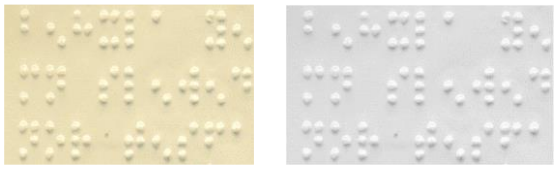


Figure3. *Braille document image (left with gray level scale and right with color)*

The distributions of noise in the Braille images are not equal. Therefore, classifying the images based on the noise level is relevant to measure the performance of the system. Due to this reason the digitized Braille images are classified into four groups (clean images, small noise, medium noise, and large noise images) based on visual criteria of the noise level on the binarization stage.

2. Image Preprocessing

Our preprocessing stage which includes adaptive histogram equalization, Gaussian and adaptive Median filtering, thresholding/binarizing, and morphological operation. This phase make the initial image more suitable for later computation.

a) Adaptive histogram equalization

It enhances the contrast of the intensity image by transforming the values of the pixels. After applying adaptive histogram equalization, the result shows that the high variation of neighborhood pixels in the image is adjusted. This greatly improves the performance of the filtering operations.

b) Gaussian Filtering

The Gaussian smoothing method is widely used not only for smoothing signals of one independent variable but also for various image processing applications. It is a very good filter for smoothing signals or images. The amount of smoothing depends on the value of the spread parameter (i.e., the standard deviation) of the Gaussian function. We have used the default value of 0.5 standard deviation, and 3x3 and 5x5 matrix kernel.

c) Adaptive Median Filtering (AMF)

Adaptive filtering has been applied widely as an advanced method compared with standard median filtering. Adaptive filter performs spatial processing to determine which pixels in an image have been affected by impulse noise. The Adaptive filter classifies pixels as noise by comparing each pixel in the image to its surrounding neighbor pixels. The noise pixels are replaced by the median pixel value of the pixels in the neighborhood that have passed the noise labeling test. For this study the Wiener adaptive filtering has been used. The results of Gaussian and adaptive median Filtering is shown in figure 4a and 4b respectively.

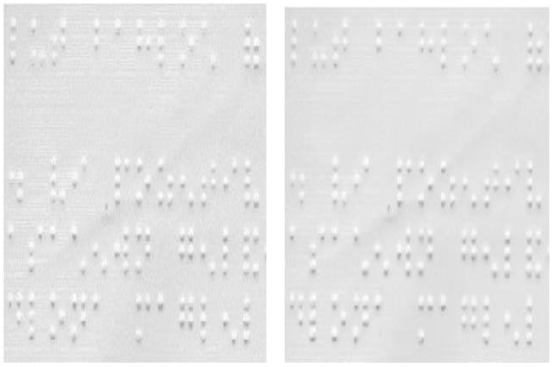


Figure 4 a) Gaussian b) AMF

The performance evaluation of the filtering operation is quantified by the PSNR (peak signal to noise ratio) calculated using formula:

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i.j) - I'(i.j)]^2 \quad \text{eq. 1}$$

and

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \quad \text{eq. 2}$$

Where, M and N are the total number of pixels in the horizontal and the vertical dimensions of the image. I and I' denote the original and filtered image, respectively. A small value for MSE means less error in the new image. Larger PSNR values signify better signal restoration because it means that the ratio of signal to noise is higher. The Gaussian and adaptive median filter is applied on the different noise level of Braille images (small, medium and high). The comparative results of AMF and Gaussian filtering with 3x3, 5x5 kernels and 0.5 standard deviation the results are presented in Table 1.

Noise Type	MSE		PSNR	
	AMF	Gaussian	AMF	Gaussian
Small	66	0	29.94	48.13
Medium	77	0	29.27	48.13
Large	119	0	27.36	48.13

Table 1. PSNR and MSE measure of Braille images

From the above results we can say that the performance of Gaussian filtering is better than adaptive median filtering in terms of PSNR and MSE. However, when we visually investigate the results of the filtered Braille image, the Gaussian filtering algorithm blurs the image, whereas the adaptive median image filtering algorithms results in no blurring effect in the image and also preserves the edge.

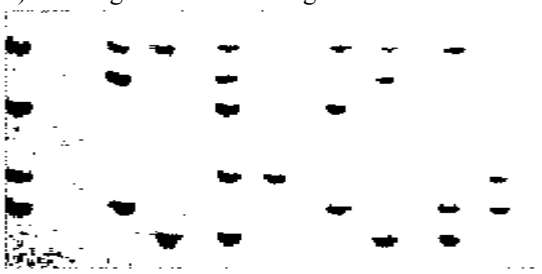
d) Binarization/Thresholding

Image binarization converts the gray scale image into binary (black and white) image aiming to distinguish text areas from background areas. For this study, we apply the Otsu's global binarization technique on the gray scale filtered image.

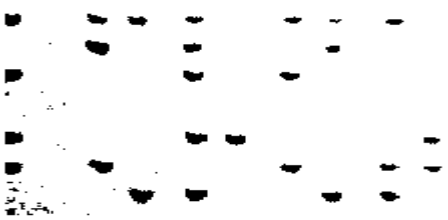
As shown in figure 5(a), some parts of the Braille image is difficult to identify the dots from the background and some of the dots also connected. Increasing or decreasing the Ostu's algorithm diminish the generated noise. So after extensive experimentation decreasing the threshold value by 0.065 from the Otsu's global threshold value generates a better Braille image, as presented in figure 5(b). This image also has some noise on the edges of the document. This is due to the fact that the edge of the Braille is the most touchable parts. To eliminate these edge noises we cut four pixels from the four sides of it. This technique generates a better result without affecting the Braille dots (see figure 5 c).

The new global thresholding technique produces a satisfactory result for clean, small and medium level noise with some addition and deletion of dots. However, for high level noise, it generates the image by deleting some of the dots and also connecting the dots at the edge of the image. Generally, thresholding the Gaussian filtered image with additional preprocessing image

A) Otsu's global thresholding



B) The new thresholding



C) Edge enhancement

Figure 5. Binarized Amharic Braille Images

performs better than thresholding the adaptive median filtering. Because thresholding the Gaussian filtered Braille images has good dot size and almost preserves all the dots for clean, small, and medium noise level images.

After the binarization, steps the scanned image has the feature that the foreground (content of the image) is represented by black color and background with white color. The result of these images shows that still there is a noise in the document. To remove this noise we analyze the pixel size of the dots in the image. Through experimentation we identify that a dot size less 20 pixels in the documents is a noise, therefore, we avoid the dots. The output of the binarization module is fed to the next level of processing, which is image segmentation.

e) Morphological operations

Morphology is a broad set of image processing operations that process images based on shapes, size, and neighborhood. The morphological operation is performed based on a comparison of the corresponding pixel in the input image with its neighbors. The implementation of morphological operations is by using MATLAB built in function 'bwmorph' majority operations. The function sets a pixel value to 1 if five or more pixels in its 3x3 neighborhood are 1 otherwise sets the pixel value to 0. The complemented result of morphological operation on the noisy Gaussian filtered image is presented in figure 6.

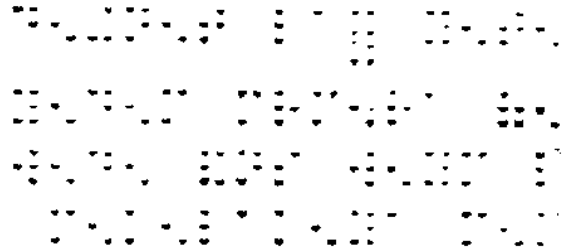


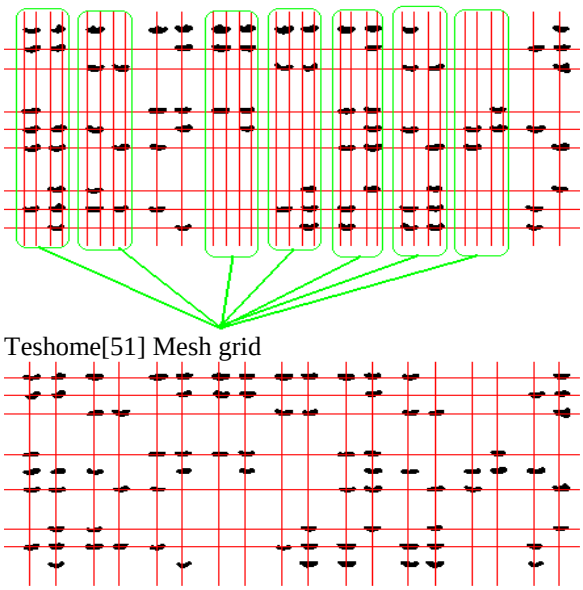
Figure 6. Results of morphological operation on the Gaussian binarized image

The result shows that the small noises in the images clearly removed. This image greatly improves the performance of the segmentation processes because the mesh grid segmentation is constructed based on the positions of pixels in the image.

3. Segmentation

Segmentation refers to the process of separation of dots from Braille image that can further be grouped into a cell. These cells further grouped into character, words of any strokes in Braille. The output of this step is used as an input to feature extraction. In the present research, mesh grid segmentation algorithms adopted by Teshome [13] is used. This algorithm is efficient in segmenting Braille image that are not connected or skewed with similar dot size The variation of the Braille dots greatly affects the construction of mesh grid (which is size dependent). Therefore, mesh grid segmentation algorithm is modified such that the mesh grid is constructed by finding the minimum and maximum dot size of the Braille dots.

The comparison of mesh grid constructed before and after modification of the algorithm is shown in figure 7. As shown in figure 7(a) the algorithm adopted by Teshome [13] constructs two mesh grids for a single dot.

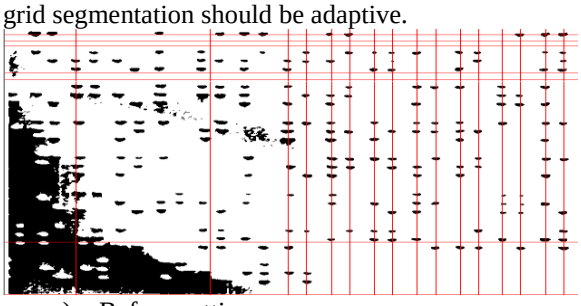


Teshome[51] Mesh grid

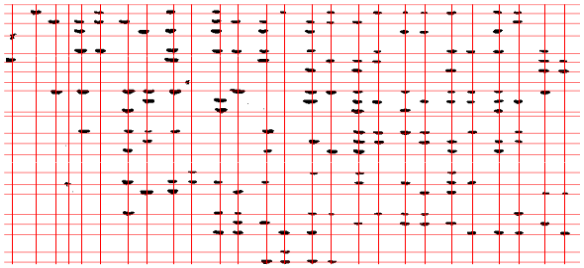
After modification

Figure 7. Segmented Braille Dots

By finding the minimum and maximum dot size the mesh grid depicted in figure 7 (b) is generated. The modification of the mesh grid is not working for all image therefore, the system to be efficient the mesh grid segmentation should be adaptive.



a) Before cutting



b) After cutting

Figure 5.13 Mesh grids for high level noise a) with large noise b) after cutting the large noise

Cutting the large noise is not a best solution because it removes the Braille dots with noise. This greatly affects the performance of the system.

4. Feature Extraction

The main function of this phase is to extract the Braille dots from the segmented image and groups them into cells [3]. The extracted features are used for testing the developed neural network model. To this end, the present study is used context based feature extraction algorithm adopted by Teshome [13] with some modification. The modification is that the previous system store the extracted feature only for the current Braille image, however the current system appends the output from the previously extracted features. The context analysis has performed to determine the status of dots in a cell. Based on the result, the cell can be recognized as Braille character alone or part of a Braille characters. This is because depending on the context, a Braille character may have one, two or three cells. The failure of constructing vertical or horizontal mesh grid highly affects the feature extraction of the Braille document images. Therefore, the success of the feature extraction is highly depends on the segmentation phase. The context based feature extraction is successful for Amharic Braille recognition. All the features of the Braille dots are extracted and represented with 6 and 12 bits that can turned on(1) or off(0) in any combination.

5. Recognition

After extracting the features of the Braille dots they have been feed to the artificial neural network. In this study, the architecture of the neural network created for the recognition is a feed-forward, supervised, multilayer perceptron (MLP) network with three (3) layers – input layer, one hidden 90 layer and an output layer. The binary representations of Braille character as input vector are created from Amharic Braille alphabet in two standards (6 bits and 12 bits) that have equal number of input nodes to the corresponding input layer of the network. The following graphical presentation gives the visual impression of the overall architecture of the recognition system starting from the Braille image input to the recognition phase.

MSE= Four training dataset is prepared with the following proportion to train the networks:

Training dataset 1: contains 34 records for basic Amharic Braille character (that use one cell representation) and the corresponding print character code.

Training dataset 2: contains 238 records for basic Amharic character (including the variant 34 *7), each character presented with 12 bit (2 cells) and the corresponding print character code.

Training dataset 3: contains 257 records for basic Amharic character (38 x 7) and punctuation marks (19), each with 12 bit (2 cells) and with the corresponding print character code.

Training dataset 4: contains 267 records for basic Amharic character (38 x 7), numerals and punctuation marks (19), each with 12 bit (2 cells) and with the corresponding print character code.

For the purpose of this project the optimal parameters used are: 0.01 learning rate, maximum of 1500 number of Epochs, and 0.0001 mean error threshold value.

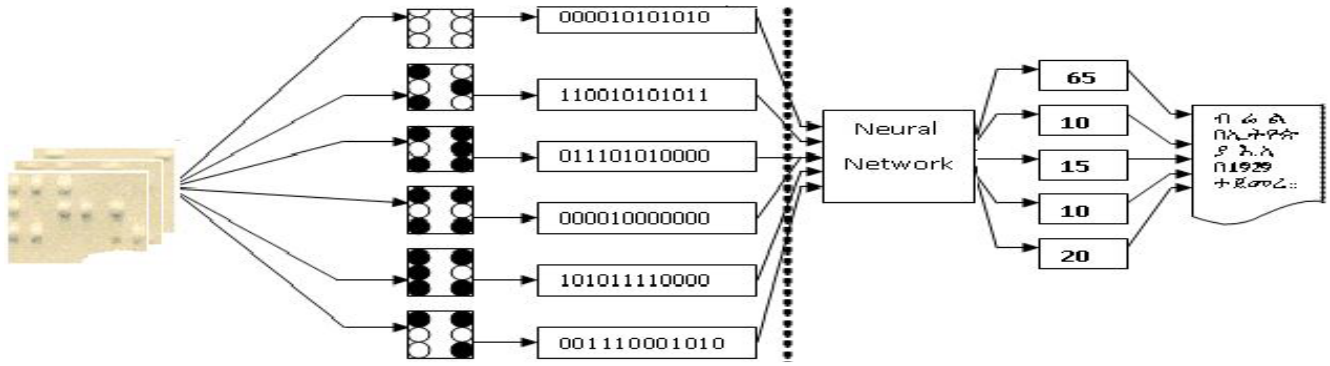


Figure 8 General Architecture of Braille Recognition (taken from Teshome[51])

Accordingly, to train the network, the four Braille character input file along with the corresponding target character file is given to the neural network. Then, it has been repeatedly trained with the original dataset for 50 iterations, which makes a total of 1500 for each character pattern in the training set.

The two artificial neural networks created for two different numbers of input nodes with four training sets are tested for accuracy of classifying new character pattern to one of the 34 and 267 characters of the Amharic language. The output of the network could range between 0 and 1 because the Amharic characters can't be represented by ASCII or the Unicode. Then, this value can be converted back into the indexed value in the range 1-34, and 1-267. The output of the two networks is presented in table 2.

Network	Performance/rate of recognition			
	Train 1 (34)	Train 2(238)	Train 3(257)	Train 4(267)
With 6 input nodes	100%	-	-	-
With 12 input nodes	-	100%	97.67%	97.75 %

Table 2. Performance of Neural Network

The above table shows that training the network with first order basic characters sets i.e. the 34 basic Amharic characters results in 100% recognition accuracy. This means the network recognizes the basic characters without any miss. However, with this network we cannot test the other training sets (i.e. training set 238, 257, and 267) because there is an input node variation.

Testing the network with 12 input nodes with the three training sets generates better results for training set 2. All basic characters with their variant characters are correctly recognized. For others (training set 3 and 4) almost the same level of performance is achieved.

To test the performance of the system, new test sets extracted from Braille document are used. The model has been tested with a set of Braille characters for each of the four Braille types.

Test set 1: contains a total of 200 characters with two cells for clean image.

Test set 2: contains a total of 200 characters with two cells for small noise level Braille image.

Test set 3: contains a total of 200 characters with two cells for medium noise level.

Test set 4: contains a total of 200 characters with two cells for high noise level.

The system is tested on the Braille images filtered with Gaussian filtering and adaptive median filtering. The performance of the neural network for the two filtering is presented in the following table.

Test Data set	Test Data size	Performance	
		Gaussian	AMF
Testset1	200	95.5%	95.5%
Testset2	200	95.5%	95.5%
Testset3	200	90.5%	87.5%
Testset4	200	65%	58.5%

Table 3 Performance rates for test dataset

The performance of the system for test set 1 (clean Braille images) and test set 2 (small noise) is the same however, accuracy decreases from test set 2 (small noise Braille) to test set 4 (high noise Braille). This indicates that the system clearly removes small noise level from the Braille image like salt and pepper, Gaussian etc. The performance of the system decreases on test set 3 is because of the addition and deletion of Braille dots in the image. The reason for high performance reduction on test set 4 is addition and deletion of Braille dots, the impact of the repetitive use, bends and highly connected Braille dots of the Braille image. Due to this, both filtering techniques are inefficient to restore the removed dots and separate the connected Braille dots.

In addition to this, the mesh grid segmentation is not handling the large dots that are found at the edge of the Braille image. In general, the decreasing performance of the system in both filtering techniques is that some characters found in the test document for which the network is not recognize in the learning process.

The performance of the Gaussian is outperforms the adaptive median filtering. This is because Gaussian filtering is better to restore dots, and also the dots have high similarity of Braille dot sizes. Therefore, for real world Braille recognition system we recommend the use of Gaussian filtering with morphological operations. There is a great variation of performance between high level noise with other (clean, small noise, and medium noise images). The reason is that the technique used to remove high connected noise with dots in the high level noisy image is cutting. This technique removes the dots that are parts of the image. Therefore, to improve the system the high level noise needs advanced noise removal techniques that clearly separate the noise with contents.

IV. CONCLUSIONS AND RECOMMENDATION

In this study an attempt has been made to recognize the real life Amharic Braille document images. Since most real life Braille document images are very poor quality, this research considers different preprocessing algorithms like interpolation, noise filtering, morphology operations, and global thresholding. The results of preprocessing stage enhance segmentation, feature extraction and the classification task.

Braille Images are very exposed to noise due to finger reading system and poor quality of the paper. To handle such different types of noise, Gaussian filtering with morphological operation is better. Gaussian filter is robust in retaining the small Braille dots and remove Gaussian noise.

Test results show that there are problems of dot connectivity in some of the Braille documents. Hence, an attempt to develop generally applicable binarization algorithm, we considered by adjusting the Otsu's global thresholding algorithms. The newly developed algorithm is tested with different intensity noise level Braille in which case a remarkable result is obtained.

Finding minimum or maximum size dots (adaptive thresholding) greatly enhance the performance of the mesh grid segmentation algorithm. Feed-forward back-propagation artificial neural network is used for classification purpose.

The Amharic OBR system is tested to evaluate its performance using different noise level Braille image. The results vary from 96.5% (for clean image) to 65% (for highly noisy image) with the type of noise level. The severity of the noise increases the performance of the system is decrease.

The most common error in Amharic Braille recognition system for medium noise level is addition and substitution where as in high level noise addition, substitution and deletion. However, in the case of clean and small noise image the main error is substitution due to the problem of neural network classifier.

Normalization techniques should be devised and integrated with present development in Amharic OBR, so that the system will be size independent.

This system uses properly scanned Braille images. The slanting of the Braille images affects the construction of the mesh grid segmentation. Therefore, incorporating skew/tilt detection and correction algorithm should be developed to enhance the performance of the system.

To enable the Amharic OBR system search for obvious errors and locate possible alternative for unrecognized words, there is a need to integrate post-processing techniques (with the help of spell checker, thesaurus, grammar, etc tools).

Since the previously adopted segmentation algorithms is not flexible enough, as expected, to make a generalized one for the recognition of different dot size, more flexible segmentation algorithm (such as adaptive mesh grid etc.) should be adopted.

Nowadays, the uses of double sided Braille are started due to the coming of Amharic Braille printer. The existence of this printer generates a huge amount of information. However, the previously developed OBR system only considers single side Braille documents. So, we propose to design an algorithm that recognizes a double sided Braille document.

Reference

1. AbdulMalik Al-Salman, et. al., "An Arabic Optical Braille Recognition System", Proceedings of the First International Conference in Information and Communication Technology & Accessibility. Hammamet, Tunisia, 2007.
2. Clutha Mackenzie, "World braille usage: A survey of efforts towards Uniformity of Braille notaion", Poor l'Oragnization des Nations Unies pour l'Education, la Science et la Culture, de l'Imprimerie, Pairs, 1953.
3. C. Ng, V. Ng and Y. Lau, "Regular feature extraction for recognition of Braille", In Proceedings of Third International Conference on Computational Intelligence and Multimedia Applications, ICCIMA'99, pp. 302-306, 1999.
4. C. Ng, V. Ng and Y. Lau, "Statistical Template Matching for translation of Braille", In Proceedings of the Spring Conference on Computer Graphics (SCCC 1999), pp. 197-200,1999.
5. Ehquierdo and M.Ghanbari, "Nonlinear Gaussian filtering approach for object segmentation", Proc.-vis. Image Signal Process, 146(3), 1999.
6. English Braille: American edition 1994. Available at: http://www.brl.org/ebae/, accessed on October 10, 2009.
7. H. Hwang and R. Haddad, "Adaptive median filters: new algorithms and results", Image Processing, 4(4):499-502, 1995.
8. Iain Murray and Andrew Pasquale, "A Portable Device for the Translation of Braille to Literary Text", 2006. Available at: http://www.cucat.org/general_accessibility/accessibility/Publications/Murray-Pasquale-Braille%20Translator%20Asset06.pdf.
9. Mahmoud Saeidi, M. Hasan Moradi and Fatemeh Sagafi, "Filtering Image Sequences Corrupted by Mixed Noise using a New Fuzzy Algorithm", Iran Telecommunication Research Center, Amirkabir University of Technology, Tehran, Iran, 2006.
10. Néstor Falcón et. al., "Image Processing Techniques for Braille Writing Recognition", EUROCAST, 2005.
11. Omar Khan Durrani K C Shet, "A New Architecture for Brailee Transcription from Optically Recognized Indian Languages", 3rd International CALIBER, 2005.
12. Ritchings R.T et. al., "Analysis of scanned Braille document", Document Analysis system, A. Dengel and A.L. Spitz (eds.), World scientific Publishing co, 1995.
13. Teshome Alemu, "Recognition of Amharic Braille Recognition", (MSc. Thesis). Addis Ababa University, Department of information science, Addis Ababa, 2009.
14. Trupti Patil, "Evaluation of Multi-core Architectures for Image processing Algorithms", Master Thesis the Graduate School of Clemson University, 2009.
15. በኢትዮጵያ የአይነስዉራን ብሔራዊ ማህበር 12ኛ ዙር ስራ አስረጻሚ ኮሚቴ የተቋቋመው የአማርኛ ፊደል አሻሻይ ኮሚቴ። የአማርኛ ፊደል አሻሻይ ኮሚቴ ዘገባ። ታህሳስ 19 እና 20 ቀን 1995 ዓ.ም. ለተጠራዉ አገር አቀፍ ጉባኤ የቀረበ። 1995
16. ጌታነህ አበበ። የአማርኛ ፊደል ከነሃሴ 24-28/1991 በሰበታ አይነሰዉራን ትምህርት ቤት አዉደ ጥናት የቀረበ። በአጸፀደ ሕፃ/ልዩና መ/ተ/ስ/ት/ዝ ዋና ክፍል። ሥ/ትም/ዝ/ክ/ም/ ኢንስቲትዩት።1984

Performance Evaluation of SSL V3.0 and Elliptic Curve Cryptography against RSA Over network communication between client and server

Dr.B.Barani Sundaram 1, Dr.N.R.Reddy 2

1Associate Professor, department of computer and information technology, Defence Engineering College, Bishoftu, Ethiopia.

2Associate Professor, Mekelle institute of technology, Mekelle, Ethiopia.

bsundar2@gmail.com,nrr26000@gmail.com

***Abstract*-Secure socket layer is a computer networking protocol for securing connection between client and server over unsecure network. It is an important technology in business sector as well as an active area of research. SSL plays very important role for SSL protects confidential information through the use of cryptography. Sensitive data is encrypted across public networks to achieve a level of confidentiality. Server and clients often use SSL for secure communications where the server must have a public and private key pair. The server uses its private key to sign messages to clients. The server sends its public key to clients so these clients can verify that the signed messages come from the server and so they can encrypt messages to the server. The server then decrypts these messages with its private key. SSL is the most widely used security protocol on the network so more efficient SSL have a significant impact on the key size and execution time. This thesis paper compares these two cryptosystems identify the significant advantage of one over the other when implemented the simulation of SSL with RSA perform lower while the winning combination remain ECC with SSL as per the metrics like key size, key generation time ,encryption and decryption time, end to end delay and average throughput in NS-2 simulator.**

Key words- **Secure socket layer, ECC, RSA NS-2**

I. INTRODUCTION

1.1 Background

Secure Socket Layer is a cryptographic protocol which has been used broadly for making secure connection between client and server. SSL protocol consists of four sub protocols: SSL Record, SSL Handshake, SSL Alert and SSL Change Cipher Spec. Each sub protocol has a clear function: SSL Handshake allows entities to establish sessions and connections, SSL Record exchanges application data between entities, SSL Alert reports on error during SSL execution, and SSL Change Cipher-Spec sets the active configuration that uses the just negotiated parameters. SSL Alert and SSL Change Cipher Spec consist of only one message that is usually exchanged during the handshake phase.

SSL is divided into two layers, with each layer using services provided by a lower layer and providing functionality to higher layers. The SSL record layer provides confidentiality, authenticity, and replay protection over a connection-oriented reliable transport protocol such as TCP. Layered above the record layer is the SSL handshake protocol, a key exchange protocol which initializes and synchronizes cryptographic state at the two endpoints. After the key-exchange protocol completes, sensitive application data can be sent via the SSL record layer.

SSL 2.0 had many security weaknesses which SSL 3.0 aims to fix. A short list of the flaws in SSL 2.0 which includes weakened authentication keys, weak MAC construction, although post-encryption seems to stop attacks. SSL 2.0 feeds padding bytes into the MAC in block cipher modes, but leaves the padding length field unauthenticated, which may potentially allow active attackers to delete bytes from the end of messages. There is a cipher suite rollback attack, where an active attacker edits the list of cipher suite preferences in the hello messages to invisibly force both endpoints to use a weaker form of encryption than they otherwise would choose; this serious flow limits SSL 2.0's strength to ―least common denominator" security when active attacks are a threat. Hence we Adopt to SSL 3.0.and tend to analyse its combination with ECC and RSA which will eliminate most of the said weakness but the performance of the combination needs a strategic simulation and evaluation.

Elliptic Curve Cryptography is emerging as an attractive public-key cryptosystem for wired or wireless environments. Compared to traditional cryptosystems like RSA, ECC offers equivalent security with smaller key sizes, which results in faster computations; lower power consumption, as well as memory and bandwidth savings. This is especially useful for devices which are typically limited in terms of their CPU, power and network connectivity. However, the true impact of any

public-key cryptosystem can only be evaluated in the context of a security protocol

1.2 Statement of the Problem

Data transmission is the greatest challenge in network between client and server .This poses great threat to sensitive data during transmission over the network. In this research we are going to combine SSLv3 along with ECC and implement it

1.3 General of Objective

The main objective of this research is to simulate the Combination of SSLv3 with ECC key exchange method to improve security for network communication between client and server.

1.3.1 Specific objective

- To study SSL V3.0 and its performance during SSL key exchange along with handshake.
- To study the ECC to combine it with SSL V3.0 to improve secure communication.
- To simulate and implement RSA based key exchange mechanism in to the Secure Socket layer protocol V 3.0.
- To simulate and implement ECC based key exchange mechanism in to the Secure Socket layer protocol V 3.0.
- To evaluate the performance of our simulation with regard to metrics like key size and Execution time.
- To report and recommend the findings.

II. .LITERATURE REVIEW

Secure socket layer is standard security protocol technology secure network communication between client and server .the researchers provides a lot of solutions based on different security protocols. Parshotam and RupinderCheema and AayushGulati, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012[4].In this paper, the authors proposed SSL which relies upon the use of dependent cryptographic functions to perform a secure connection. The first function is the authentication function which facilitates the client to identify the server and vice versa. They have used several other functions such as encryption and integrity for the improvement of security. The most common cryptographic algorithm used for ensuring security is RSA. It still has got several security breaches that need to be dealt with. An improvement over this has been implemented in this Research. They have proposed a modification of RSA that switches from the domain of integers to the domain of bit stuffing to be applied to the first function of SSL that would improve secure communication. The introduction of bit stuffing will complicate the access to the message even after getting the access to the private key. So, it will enhance the security which is the inevitable requirement for the design of cryptographic protocols for secure communication.

Ahmed, Defense University, College of Engineering Department: Computer and Information Technology" June 2013[5]. In this paper, the authors proposed theSSH protocol for secure communication between client and server used to remote login network services over an insecure network. SSH is intended to run over a reliable transport protocol, such as TCP. There are two versions of SSH, imaginatively called SSH1 and SSH2. Use of SSH1 is deprecated because of some security problems. SSH2 has been separated into modules and consists of three protocols working together:

SSH Transport Layer Protocol (SSH-TRANS)
SSH Authentication Protocol (SSH-AUTH)
SSH Connection Protocol (SSH-CONN)

SSH is designed to be modular and extensible. All of the core protocols define abstract services they provide and requirements they must meet, but allow multiple mechanisms for doing so, as well as a way of easily adding new mechanisms. All the critical parameters of an SSH connection is negotiable, including the methods and algorithms used in:

- Session key exchange
- Server authentication
- Data privacy and integrity
- User authentication
- Data compression

SSH-TRANS is the fundamental building block, providing the initial connection, record protocol, server authentication, and basic encryption and integrity services. After establishing an SSH-TRANS connection, the client has a single, secure, full-duplex byte stream to an authenticated peer.

Ankita Nag, Vinay Kumar Jain, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-4 Issue-4, September 2014[6]. In this paper, the authors proposed theRSA has stronger security than single key cryptography. RSA has a pair of key private key and public key. Sender sends the message encrypting it with the public key of receiver. Receiver receives the message by decrypting it with its private key. RSA provides authentication and integrity. So it is used in SSL for key exchange. At present 512 bit is considered insecure after the implementation of General Field Sieve Number. So the idea of bit stuffing is introduced. RSA is bit stuffed after encryption that means a random number is

appended to the cipher text and sent. At receiver, stuffed bit that is that random number is removed and then the cipher text is decrypted. Bit stuffing is suggested as a logic or measure to be used instead of increasing the number of bits in RSA. Since larger bit numbers will require more time and effort for calculation, bit stuffing will save time and effort. In this paper, this idea is implemented in hardware. Same security as with larger bit number say 1024 can be get in almost same time with lesser bit numbers say 512 bits with lesser band width requirement.

RSA Algorithm is generally used in SSL for key exchange between client and server or sender and receiver not for message exchange since it is 1000 times slower than symmetric key

Poonam Ashok Kakade, International Journal of Trend in Research and Development, Volume 2(5), October 2015 [7]. In this paper, the authors proposed the secure socket layer is a security protocol, that provides a secure channel between a client and a server at the transport communicating parties over the Internet's protocol is designed to authenticate the server and the client.

Swadeep Singh, AnupriyaGarg, AnshulSachdeva, International Journal of Computer Science and Communication Engineering IJCSCE Special issue on "Recent Advances in Engineering & Technology" NCRAET-2013[8]. In this paper the authors proposed the RSA is the most popular public-key cryptosystem today but long- term trends such as the proliferation of smaller, simpler devices and increasing security needs will make continued reliance on RSA more challenging over time. Hence Elliptic Curve Cryptography (ECC) is a suitable alternative. This paper focuses on performance attribute of public key cryptosystems. The algorithms studied and compared are RSA, ECC. They have implemented these algorithms in Java in order to perform software tests so that we may gain insight into the relative performance of each algorithm and its associative parameters. Software based tests are performed to yield an overall analysis of key generation, message encryption and decryption. Implementations are in Java and executable in the Windows environment. Each algorithm is tested for key generation and encryption/decryption of ordinary but large files

III. MATERIALS AND METHODOLOGY

3.1Tools and Softwares Used:

- Ubuntu 12.04 is used as Operating system.
- Network Simulation (NS2) is used to simulate the proposed system.
- C++ and C programming language are used.
- Tcl, Awk, OTcl script languages.
- Crypto tool 1.431 Beta 6b [Vs 2008].

3.2 Simulation Tool

3.2.1 Network Simulation2 (NS2) Tool

All simulations have been carried out using the NS simulator program version 2.35 under Ubuntu operating system. NS2 is a discrete-event driven object-oriented network simulator and it is open source simulator software used by a lot of institutes and researchers.NS2 has been written in two languages, Object oriented variant of Tool Command Language (OTCL) and object oriented language C++. While the C++ defined the internal mechanism (backend) of the simulation objects, the OTcl sets up simulation by assembling and configuring the objects as well as scheduling discrete events (frontend). NS2 creates two main analysis reports simultaneously [14]. One is NAM (Network Animator) object that shows the visual animation of the simulation. The other is the trace object that consists of the behavior of all objects in the simulation. Former is .nam file used by NAM software that comes along with NS. Latter is a —.tr‖ file that includes all simulation traces in the text format.

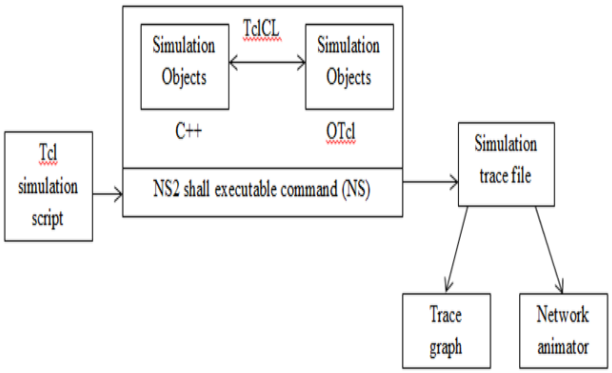


Fig 3.1 Basic Architecture of NS2

3.2.2 CrypTool-1 (CT-1)

Is a free, open-source Windows program for cryptography and cryptanalysis? It is available in 5 Languages and the most wide-spreaded e-learning software of its kind. It supports both contemporary teaching methods at schools and universities as well as awareness training for employees and civil servants.

Originally designed as an internal business application for information security training, CrypTool- 1 has since developed into an important open-source project in the field of cryptology and IT security awareness. CrypTool 1 is written in C++[15]

The current version of CrypTool 1 offers the following highlights:

- Visualization of several algorithms (Caesar,

Enigma, RSA, Diffie-Hellman, digital signatures, AES, etc.)

- Numerous classic and modern cryptographic algorithms (encryption and decryption, key generation, secure passwords, authentication, secure protocols, etc.)
- Cryptanalysis of several algorithms (Vigenère, RSA, AES, etc.)
- Crypt analytical measurement methods (entropy, n-grams, autocorrelation, etc.)
- Related auxiliary methods (primality tests, factorization, base64 encoding, etc.)

3.3 Design and Implementation of SSL in Network Simulator

- SSL v3.0 with RSA is simulated in NS2 simulator
- SSL v3.0 with ECC is simulated in NS2 simulator
- Simulated system is evaluated for performance Against key size and Execution time
- Findings are tabulated, and graphed.

3.4. IMPLEMENTATION in Cryptool-1

Create any text file in the format of .txt

- Select the size of text file
- Then open PKI then chose generate/import key
- Finally chose one by one ECC and RSA generation keys and evaluated the result in millisecond After creating the key encrypt and decrypt the text files then evaluate the result in millisecond
-

3.4.2 IMPLEMENTATION OF SSL IN NS-2

To implement the new SSL protocol, we start by patching SSL protocols in NS-2 ssl directory and named the directory as ―ssl‖ (.h and .cc files are modified). All the files in the SSL directory are modified with SSL such as ssl.cc, ssl.h, .The new protocol will use the same SSL packets so we have changed the names of sslrsa.h code and sslecc.h code. By creating all this we have designed SSLwith RSA and SSL with ECC protocols to send packets with each other. To integrate the sslrsa.h and sslrsa.cc protocol to the NS2 and also integrate the sslecc.h and sslecc.cc protocol to the NS-2, four common files has to be modified. The new packet name has to register to the packet.h. Of course, the ns-packet.tcl has to be modified so that the new class is complied. At the TCL layer, the new packet must be declared by adding the name and default packet size value to the ns-default.tcl file. Finally, we have to make an entry for the new packet in the makefile file.

- Open the ~ns-2.35/common/packet.h file there will be two changes
- Change one

Static const packet_t PT_ Security_packet =128;

- Another change

name_[PT_SECURITY_PACKET]="
Security_packet ";

- The next file to be modified Open the file ~ns-2.35/tcl/lib/ns-default.tcl, and paste the following Agent/Security_packet set packetSize_ 0
- The next file to be modified open the file ~ns-2.35/tcl/lib/ns-packet.tcl, Input the following line
  Security_packet;
- The last file modified is the makefile.in in the root directory of NS-2.35. This file is modified for creating the object files for the C++ coded files. After all the implementations are ready, we have to recompile NS-2 again to create the object files the objects are created step by step
- first ssl/sslrsa.o
- after finishing , delete and create
- ssl/sslecc.o
- OBJ_CC = ssl/sslrsa.o
  ssl/sslecc.o

 RSA ,ECC with SSL v.3 are simulated in NS2 with following topology

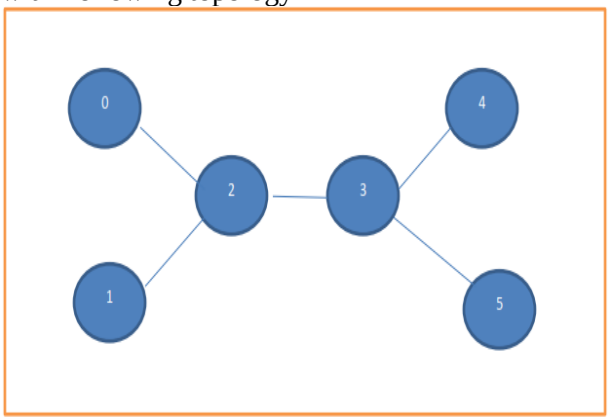


Fig 4.2 Proposed Topology

IV. EXPERIMENTS AND EVALUATION

In this section, we evaluate the impact of SSL protocol with in RSA and ECC algorithms. To evaluate these models, key generation, encryption, decryption packet delivery ratio, end to end delay is used. And we provide a detailed analysis that obtained from the simulation results.

4.1 MODELING OF NETWORK

The experiments are simulated using ns-2. The size ofthe network is specified by manually.

4.1.1 Simulation Parameters

In our simulation we used the Secure_packet connection. SS with RSA and SSL with ECC protocols are implemented then the Secure_packet will close the connection.

We simulated a network with 6 nodes and create a Security_packet connection between two nodes andpacket size is chosen to be 128 bytes long and the packets are generated at an interval of .05s.

A brief summary of the simulation parameters are listed in Table 4.1.

Parameter	Value
Simulator	NS-2(version 2.35)
Connection type	Security_packet
protocol	SSL
Number of node	6
Packet size	128 bytes

Table 4.1 Simulation Environment

4.2 Performance metrics

The performance of SSL protocol with in RSA and ECC are measured by certain quantitative metrics. The performance metrics chosen for the evaluation of SSL with RSA and SSL with ECC are packet delivery ratio, throughput and end to end delay.

The performance metrics chose for evaluation ECC and RSA algorithms interims of key size and execution time.

End to End Delay: Finishing time of data packets subtract by starting time of data packets transmitted by all nodes. This performance metric will give us an idea of how well the protocol is performing in delay.

End to End Delay= (end time of packet- starting time of packet)……..... (1)

Average throughput: it is the ratio of total amount of data which reaches the receiver from the sender to the time it takes for the receiver to receive the last packet.

Average throughput = (number of bytes received x 8 /simulation time x 1000) Kbps…… (2)

4.3 Simulation scenario

We are mainly focusing on two simulation scenarios, the operation of SSL with RSA and SSL with ECC protocols.

4.3.1 First Scenario

In these experiments SSL protocol with RSA algorithm. The network size is 6 nodes and is manually distributed. Security_packet connections are established between the 0(sending node) and 5 (receiving node) and vise versa

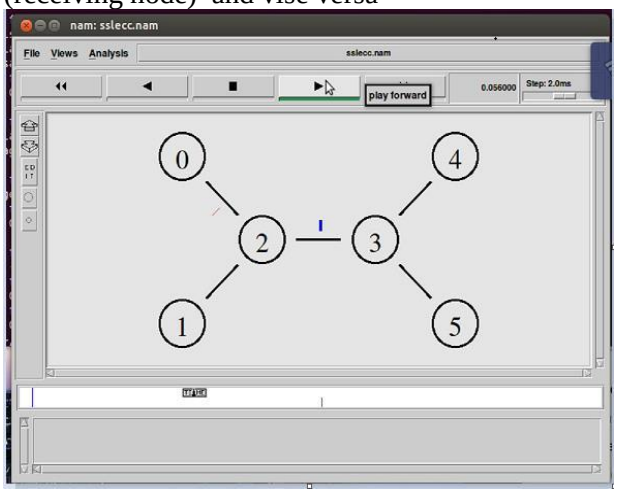


Figure 4.2 SSL protocol with RSA algorithm

4.3.2 Second Scenario

In the second simulation scenario SSL protocol with ECC algorithm. The network size is 6 nodes and is manually distributed. Security_packet connections are established between the 0 node sending and 5 node receiving nodes and vise versa.

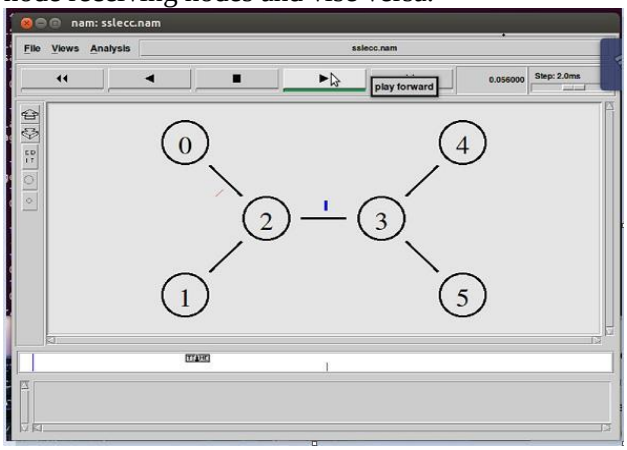


Figure 4.3 SSL protocol with ECC algorithm

V. RESULTS AND DISCUSSION

This focuses on result and its analysis based on the simulation performed in NS-2 & Cryptool-1. Our simulated results are provided in the Figures below gives the variation in network. To evaluate the simulation result, we considered the performance metrics of key generation, execution time of encryption and decryption, end to end delay and average throughput.

5.1 Key Size Comparison (in bits) for Equivalent Security Level:

ECC reduces the key size and key generation; So ECC can provide strong security by using lower key size and time key generation than any other asymmetric key encryption algorithm RSA.

- **Key generation (ms) ECC and RSA**

Security Bits	RSA	ECC	Key generation (ms) ECC	Key generation (ms) RSA
80	512	192	0.042	0.012
112	768	196	0.147	0.014
128	1024	239	0.444	0.027
192	2048	256	0.388	0.270

Table 4.1a

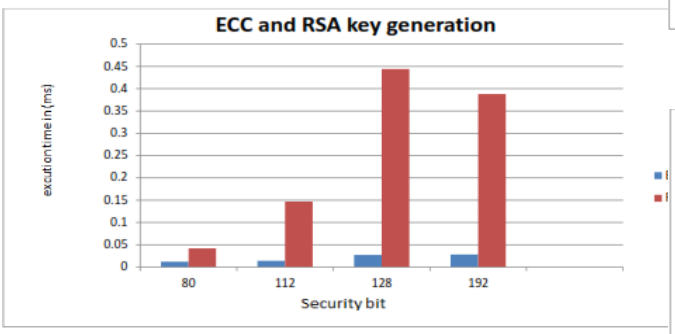


Figure 5.3 key generation Vs key size for ECC and RSA

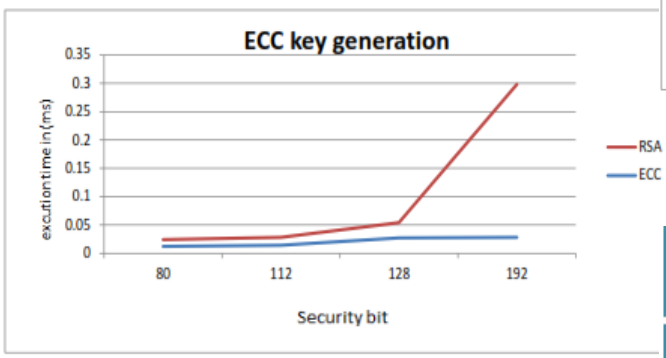


Figure 5.4 key generation Vs key size for RSA and ECC of graph

- **Encryption of ECC and RSA**

Security Bits	RSA	ECC	Time Encryption (ms) RSA	Encryption (ms) ECC
80	512	192	6.644	200.57
112	768	196	10.669	202.443
128	1024	239	11.252	204.610
192	2048	256	11.930	202.282

Tab 5.2a

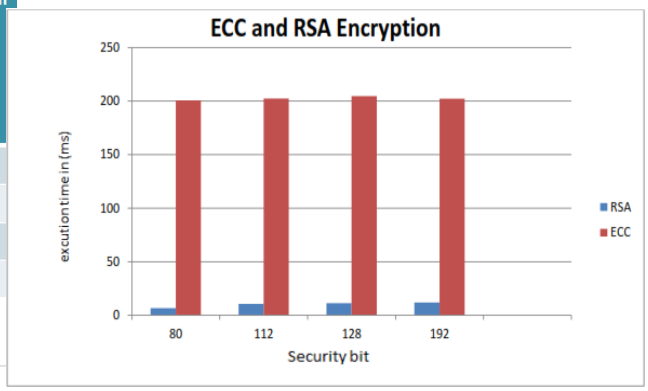


Figure 5.5 Encryption of ECC and RSA

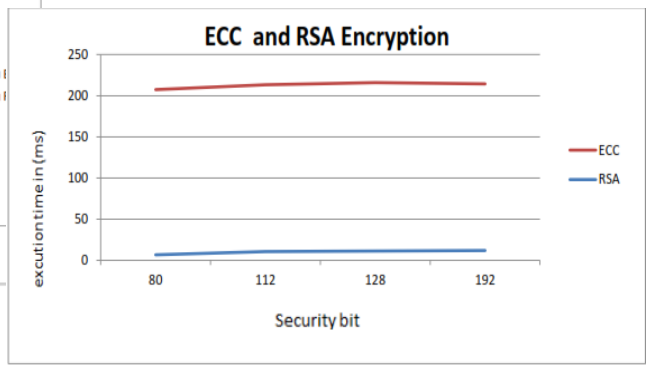


Figure 5.6 Encryption of ECC and RSA of graph

- **Decryption of ECC and RSA**

Security Bits	RSA	ECC	Time Decryption (ms) ECC	Time Decryption (ms) RSA
80	512	192	1.198	94.370
112	768	196	1.136	173.900
128	1024	239	1.177	262.390
192	2048	256	1.179	11.930

Tab 5.3a

5.2 Encryption and Decryption using Equivalent Security Level:

From the below tables it is clears that asymmetric key encryption algorithm ECC uses the lowest key size for encryption but high execution time uses for encryption but RSA key size larger than ECC but used less execution time to encryption. For decryption ECC is best preferable than RSA.

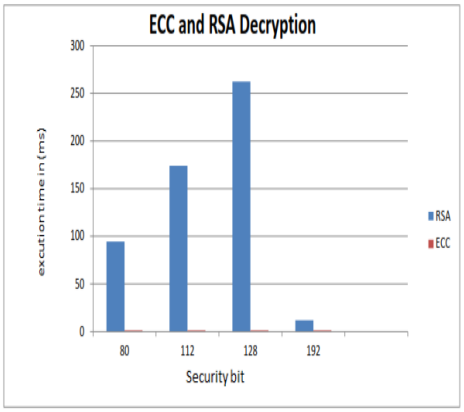


Figure 5.7 Decryption of ECC and RSA

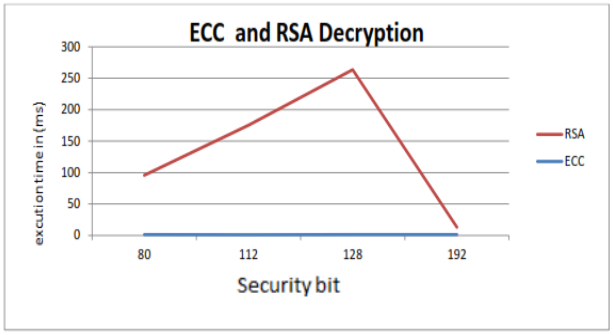


Figure 5.8 Decryption of ECC and RSA of graph

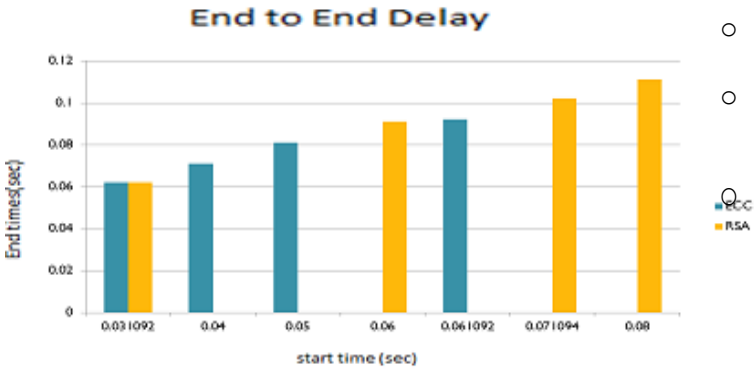


Figure 5.9 End To End Delay

- AVERAGE THROUGHPUT = (NUMBER OF BYTES RECEIVED X 8 /SIMULATION TIME X 1000) KBPS
- BASED ON RSA WE GET THE RESULT=0.0004096MBYTE/SEC.
- BASED ON ECC WE GET THE RESULT=0.002048MBYTE/SEC

5.4 DISCUSSION:

- From the above Table 1.1a result to generate key size of ECC better than RSA measured in millisecond.
- From the above Table 1.2a result key size of ECC better but it takes much time than RSA for 23kb text data.
- From the above Table 1.3a result key size of ECC small and also it takes minimum time to Decrypt than RSA for 23kb text data
- Based on the simulation result, performance of ECC algorithm is better than RSA algorithm.in key generation and Decryption
- We also observed that ECC is more secured than RSA with SSL V 3.0 protocol because of Small size, high security and also better average throughput.

VI. CONCLUSION

- Security in the network seemed to be the most interesting and complex yet the most researched area. The idea of this Research was to achieve secure communication between two nodes in the network while maximizing the security and minimize key size of transmit ion between sender and receiver.
- Elliptic curve fits in perfectly with the requirement by providing smaller key.
- To evaluate the performance of ECC, a comparison with a widely popular RSA security scheme was done.
- When RSA and ECC are compared, the overall performance of ECC was found to be better and it requires a fairly less key size to attain the similar security level as RSA. The SSL protocol used with both ECC and RSA schemes under network. The experimentation used five evaluation parameters: End to end delay, average throughput, key generation, encryption and decryption. ECC scheme in general had a better performance than RSA due to smaller key lengths

REFERENCES

[1] Parshotam , Rupinder Cheema and Aayush Gulati —Department of Computer Engineering, Lovely Professional University‖,‖ Improving the Secure socket layer by modifying the RSA Algorithm‖ June 2012.

[2] Ahemed Abdella ,‖ Secure Remote Desktop Access Using SSH and Elliptic curve cryptography on a PKI‖,‖ Defense University, College of

Engineering Department: Computer and Information Technology‖ ,June 2013.
[3] Poonam Ashok Kakade,‖ Trusting SSL for Unsecure Internet‖, October 2015.
[4] Mohammed A. Alnatheer, ―Secure Socket Layer (SSL) Impact on Web Server Performance‖, September 2014.
[5] Mr. Pradeep Kumar Panwar and Mr. Devendra Kumar ―Security through SSL‖ December 2012

[9] NIST, \Recommended Elliptic Curves for Federal Government Use", July 1999, see http://csrc.nist.gov/csrc/fedstandards.html
[10] William Stallings, Cryptography and Network Security, Principles and Practice. ed., Prentice Hall, New Jersey, 2003
[11] Aleksandar Jurisic and Alfred Menezes, ―Elliptic Curves and Cryptography‖, Dr. Dobb's Journal, April 1997.
[12]. Borst, ―Public key cryptosystems using elliptic curves‖, Master's thesis, Eindhoven University of Technology, Feb. 1997. http://citeseer.nj.nec.com/borst97public.html
[13] Mugino Saeki, ―Elliptic curve cryptosystems‖, M.Sc. thesis, School of Computer Science, McGill University, 1996. http://citeseer.nj.nec.com/saeki97elliptic.html

[6] Wade Trappe, Lawrence C. Washington, ―Introduction to Cryptography with Coding Theory 2nd Edition‖.Pearson Education, ISBN 81-317-1476-4.
[7] Ali Hilal Al-Bayatti," Security Management for Mobile Ad hoc Network of Networks (MANoN) ",De Montfort University, February 2009.
[8] Stefan Katzenbeisser. Recent Advances in RSA Cryptography. Kluwer Academic Publishers, 2001

African Buffalo Optimization based Efficient Key Management in Categorized Sensor Networks

Dr.J.R.Arunkumar[1], Dr.M.Sundarrajan[2], Dr.R.Anusuya[3], Mr. KibrebAdane[4]
[1, 3, 4] Department of Computer Science and IT, [2] Department of Electrical and Computer Engineering
Arbaminch Institute of Technology, Arbaminch University, Arbaminch, Ethiopia.
[1]arunnote@yahoo.com , [1]arun.kumar@amu.edu.et, [2]sundar.rajan@amu.edu.et ,[3]anusuya.ramasamy@amu.edu.et

Abstract **— In recent days, Wireless Sensor Network (WSN) can be used to monitor the circumstances of various movable objects and several processes such as friendly forces monitoring, biological attack detection, fire detection and so on, its applications are extended. To make all these applications reliable and secure, it is necessary to use cryptographic processes. The security of the whole network is depends on the strength of a generated key, the algorithm decides the size of the key handling and processing. In this research, a secure efficient key management scheme is proposed with the help of African Buffalo Optimization (ABO) for wireless sensor networks. The evaluation of this method is made with the objective to improve the security strength and reduce the cost of resource. The traditional methods called genetic algorithm and an evolutionary algorithm named as Particle Swarm Optimization (PSO) is compared to verify the proposed scheme**.

Keywords**- Wireless Sensor Network, Key Generation, Particle Swarm Optimization, African Buffalo Optimization, Genetic algorithm**.

I. INTRODUCTION

In sensor network, the key management is one of the core security protocols for several real time applications. Normally, many researches were focused on some constrained resource such as limited battery power and processing capabilities. Apart from these constraints security is one of the main factor that decides the data transmission between nodes. Data transmission from one node to another is depends upon the electronic security. For any legal transmission of information it is necessary to encrypt the data. For this process the cryptography is used to convert the original information into unreadable format and while retrieving the data this encrypted data is again converted into original messages. Some of the security issues in sensor nodes are stated as, the radio links are insecure and sensor nodes are not temper resistant (i.e., attacker obtains all security information).

Cryptographic methodology and cryptographic key are the two factors deciding cryptography. The algorithm decides the encryption and decryption process with the key as a parameter used by the function. The total security strength is made by the algorithm with several objectives named as key freshness, key authentication and key integrity. The mathematical process is depends upon the selection of algorithm, to provide authentication and confidentiality. The discrete logarithmic problem stated by Stinson (2005) is implemented with two prime numbers. Likewise other key generation methods include the derivation of a key from another key, these process is made by deriving the key from its password. Many traditional schemes focused on key pre-distribution schemes to make robustness in design reduce the memory requirement. In existing methods the Diffie-Hellman (Rescorla 1999) method are used for two layered key management and dynamic key update protocol. These process never suit sensors with large key sizes. The limitations of these process is unable to access the large keys, increases overload, security is less and induce key escrow problem.

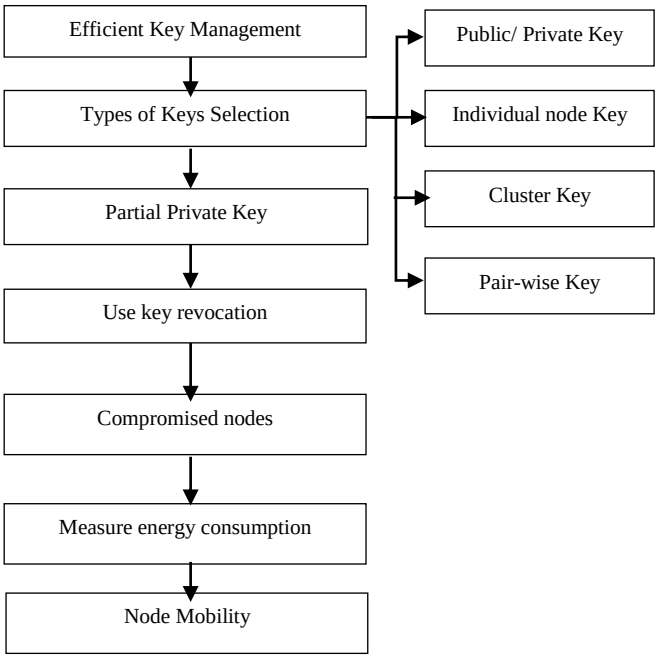


Fig. 1. Key management flow diagram

The evolutionary algorithm (EA) considered in this research is African Buffalo Optimization (ABO), it is an optimization technique established on intelligence of African buffalos and its movement. It is effective for tracking and identification of green pastures. The

solution search is based on the sound made by each buffalos. The previous implementation of ABO made in travelling sales man problem. Mainly this optimization is selected for solving several problems because of its faster speed.

This section discussed about some main problems and traditional key management scheme. For further implementation and improvement in key management is the main objective of this work. Hence, a detail review and background study is made in Section 2. Section 3 provides the exact problem which is identified from the survey. Section 4 states research methodology that contains traditional genetic algorithm, particle swarm optimization and the proposed African buffalo optimization for improving key management. To verify the experimental results, section 5 provides detail comparison among all methods. Finally, the conclusion is made in section 6 with the future directions.

II. LITERATURE SURVEY

In real time data processing environments the sensor network occupies more processing units and security issues. Lee et al., (2004) presented a key management scheme that satisfies both operational and security requirements of distributed sensor networks. They have made a data aggregation for energy efficiency and watchdog mechanism for intrusion monitoring as an additional feature. They designed this scheme for applying in variety of sensor network protocols.

In order to achieve security in sensor networks, the messages must encrypt among sensor nodes. For encryption the keys must processed by communicating nodes. Due to several resource constraints, achieving such key agreement in networks is nontrivial. Some of the traditional pre-distribution of secret keys for all pairs of nodes is not suitable because the memory is large if the network is large. The assumption is made for random key pre-distribution schemes because there is no deployment knowledge is available. To solve these factors, Du et al., (2004) presented a key management scheme by deployment knowledge. They arranged the performance metrics such as connectivity, memory usage, and network resilience.

Sarkar and Mandal (2012) proposed a Public Key Cryptography (PKC) using swarm as solitary has been carried for wireless communication, it is tested with the with the deterministic approach. Hussein et al., (2010) observed that encryption based on chaotic map that has been derived from a simplified model of Swarm Intelligence (SI). They have been proposed along with the analysis of the possibility of using SI in the field of image cryptography. To increase the robustness of the system the Swarm Intelligence Chaotic Map (SICM) is used.

TABLE I
COMPARISON OF KEY MANAGEMENT SCHEMES

Sl. No	Citation	Methodology	Advantages	Role and applications
1	Chen and Chen (2014)	Secure routing solution based on LEACH protocol	It improved the survivability of node more efficiently in a harsh sensor network environment	System security is integrated into sensor node. Clusters are changed dynamically and periodically according to node mobility.
2	Celozzi et al., (2013)	Enhanced variant of the LEAP+ protocol	Decreases the key setup time by reducing the number of packets exchanged	To improve the security of communications and implement in hardware to verify the number of security threats
3	Wan et al., (2015)	Public-Key-Based Data Integrity Protection	Reduced the signing cost, PKDIP can even be more efficient than hash-function-based schemes.	It introduces the “Montgomery Modular Multiplication (MontMM)” technique to current public-key-based signing algorithm for wireless Image Sensors
4	Rahman and Sampalli (2015)	Efficient Pairwise and Group Key Management Protocol	It is efficient in terms of computation, communication and energy overhead.	It is designed to support both pairwise and group-wise key management
5	Latiff et al., (2016)	Hybrid algorithm based on Backtracking Search Optimization Algorithm (BSA) and K-Means	It is able to deliver more data to the base station and extends the network lifetime.	Heuristic algorithm is required to produce good clustering in sensor networks.

Sreelaja and Pai (2008) made a Ant Colony Key Generation Algorithm (AKGA) based key generation to perform encryption of data in cellular network. Sreelaja and Vijayalakshmi presented an encryption method of plaintext using a stream cipher method with single AKGA. The keys generation for encryption is made by swarm intelligence approach. In order to reduce the number of keys, an Ant Colony Optimization key Generation Algorithm (AKGA) is used. Khan et al., (2013) made a Data Encryption Standard (DES) by Ant-Crypto and Binary Ant Colony Optimization (BACO). They made two metrics such as the ratio of the optimum keys in all solution and the number of success bits that are matched with the original key. Zhu et al., (2006) described a hierarchical key management method called as Lightweight Extensible Authentication Protocol (LEAP) to meet the different security requirements. They establishment is of four types of keys for each sensor node, such as individual key, group key , cluster key and pair key.

III. PROBLEM STATEMENT

From the survey it is noticed that many clustering routing protocols are proposed with security issues. The problem is stated as while encrypting the data or a message there is a need for key, the key helps to get back the original data at receiving unit. The problem occurred during encryption is key updating and clustering head selection. To optimize the clustering network, several optimization algorithms are made. Since, the key management is one of the important way to protect the clustering safety. Some of the traditional algorithms such as Genetic Algorithms (GA), improved genetic algorithm are used to solve the discrete optimization problems presented in the network based on the fitness value. The evolutionary techniques is one of the alternative method for traditional techniques, because traditional methods are in high nonlinearity and low autocorrelation. In this research, the security issues and network success rates are considered and analyzed. To solve these issues a new inspiration made from the behavior of African buffalo is considered.

IV. RESEARCH METHODOLOGY

Odili et al., (2016) presented a new optimization technique called the African Buffalo Optimization. African buffalos are a wild species of domestic cattle, it is always in mobile tracking that finds the rainy seasons in different parts of Africa. Based on the two basic sounds in search of solutions, the optimization is made. It is robust, effective, user-friendly, efficient and easy to implement. ABO try to solve the problem of pre-mature convergence by ensuring the location of each buffalo which is regularly updated in relation to the particular buffalo's best previous location. The current location of the best buffalo's location is not improved in a number of iterations, the entire herd (present location) is re-initialized. The best fitness is based on tracking the search space and tapping into the experience of other buffalos. The algorithmic steps used in key management is shown below,

Step1. Objective function $f(x)$ $x=(x_1, x_2, \ldots\ldots x_n)^T$

Step2. Initialization: randomly place buffalos to nodes at the solution space;

Step3. Update the buffalos fitness values by following equation

$w_{\cdot k+1} = w_{\cdot k} + lpr_1 (bg_{maxt.k} - m_{\cdot k}) + lpr_2 (bp_{max.k} - m_{\cdot k})$

Where $w_{\cdot k}$ and $m_{\cdot k}$ represents the exploration and exploitation moves respectively of the k^{th} buffalo (k=1, 2…N); lp1 and lp_2 are learning factors; r_1 and r_2 are random numbers between [0, 1];

$bg_{max\ is}$ the herd's best fitness and $bp_{max,}$ the individual buffalo's best

Step4. Update the location of buffalo k in relation to $bp_{max.k}$ and $bg_{max.k}$ using $m_{\cdot k}+1 = \lambda (w_{\cdot k} + m_{\cdot k})$. Where 'λ' is a unit of time.

Step5. Check bg_{max} is updating or not. If yes, go to 6. else, go to 2

Step6. If the stopping criteria is not met, go back to algorithm step 3

Step7. Output best solution.

The African buffalo optimization is modeled as three characteristics, the sound named as "maaa" of buffalo is indicated as (k=1,2,3,..,n) represented by m.k, next the sound "waaa" is represented by w.k. The initial key generation is based on the node to update the keying process.

It is directly depend on buffalos fitness. A key stream is generated based on the random or pseudorandom characters that are combined with a plaintext message or with an image to produce an encrypted message. Normally, the character in the key stream is added, subtracted or XORed with a character in the plaintext to produce the cipher text with arithmetic. The initial process is selection of bits, here 32 bit is considered, the each bit (buffalo) is validated by their fitness value. The bg_{max} is used here to measure the threshold value with respect to the African buffalo's optimization. The key generation process using ABO is represented in a flow chart as shown in the steps. The mobility is noted by the sensor nodes with respect to the buffalo's fitness value. The w.k represents the initial node bit length. A sample key obtained from the initial step based on iterations is shown in Fig. 2.

32 bit length	0010 0101 0110 0100 1010 0111 0111 0001

Fig. 2. Sample Key of 32 bit length

Key generation is the most important factor, hence, in this work ABO is used in the key generation process where key selection is depends upon the fitness function. Based on the iteration, the key has its highest fitness value.

The selection value is selected with the threshold value to compare the process. The key must be selected as unique and non-repeating. The encryption is made with the selected key and are highly encrypted because of more randomness of key.

TABLE II
PARAMETER SETTING

Types	Parameters		
GA	β=2.0	200 Nodes	ρ=3.56
PSO	p_i=1.5	150 Nodes	$t_{initial}$=0.65
ABO	m.k=1.0	150 Nodes	Bg $_{max}$=0.61

The main advantages of ABO based on key generation is more accurate and the session key can be established directly between any two (buffalos) sensor nodes. In GA the initial process is started with an initial population of 32 bits chromosomes which are generated randomly. This population is stored in an array, that contains maximum population, initPop[MAX_POPULATION][32], in this process each cell consists of binary values 0 or 1. The maximum nodes considered here is 200, MAX_POPULATION. The size of chromosome cell is equivalent to the length of key intended to be generated.

The genetic algorithm flow is represented by initialization, threshold check, followed by iterative application of operators. For processing and selection, the following format is used, the size of array, final population [] meets MAX_POPULATION limit. Parent selection, crossover and mutation is the basic process of GA. Finally, the fitness value is calculated for each chromosome. For PSO process the algorithm parameters such as Swarm_Size, Max_Iter, c1, c2, w, V_MAX are considered. First step in PSO is initialization of swarm of particles, updating the global best position followed by initializing the velocity of particles. Each particle in swarm is initialized with a random seed velocity which is bounded by the intervals V_MAX and V_MIN. by updating the particles velocity the key generation I selected. Finally, the position is updated and find the best key.

$$x[i] = x[i] + v[i]$$

where, x[i] is position of ith particle and v[i] is velocity of ith particle. After updating the threshold value the fitness function calculates the fitness of key stored at a particular location in key domain. The sequence of 0s and 1s provides the best randomness among all process.

V. EXPERIMENTAL RESULTS

The performance of the proposed African buffalo optimization algorithm is compared with the genetic algorithm and particle swarm optimization. It is implemented in network simulator, to find the ratio of the optimum keys in all solution and the bits in guessed key that are matched with the original key, these are considered as the performance metrics.

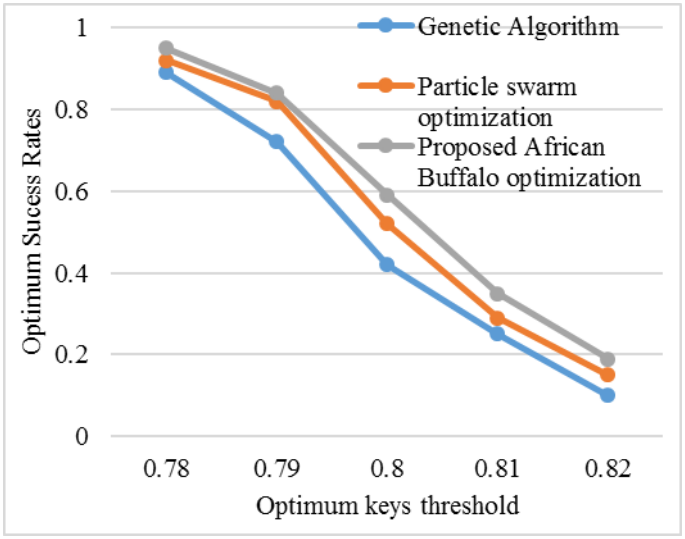


Fig. 3. Comparison of GA, PSO and ABO based on optimum success rate

The proposed work implementation is made to find Public Key Cryptography with ABO, PSO and GA algorithm. The proposed method is a suitable match for key generation with high fitness for achieving quality cryptography. This graphical representation shows that the generated keys using ABO are unique and more secure for encryption of data.

The Fig. 3 represents the comparison of traditional GA, PSO and proposed ABO methods with the optimum key generation. The success rate varied according to the iteration. If the threshold value is equal or near to unique then the original key is matched (i.e., Key is matched).

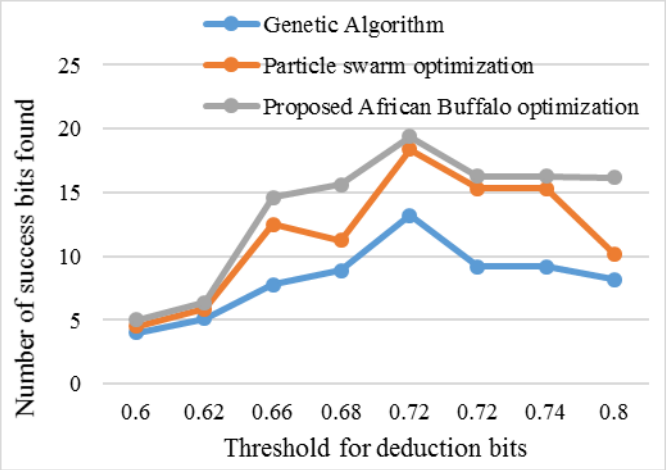


Fig. 4. Comparison of GA, PSO and ABO based on success bits

Fig. 4 shows the comparison of deduced bits that gives the number of success bits. In minimum threshold value, all the three optimization algorithms are equal, if the threshold varies then the success bits increased

The experimental results shows that it has strong adaptability and the performance factor is improved based on the threshold value. If iteration is found then the bit selection is varied and gives the success bits. It helps in each encryption and decryption process. Finally, data confidentiality and data integrity is achieved due to optimal nature of buffalo's strategy.

VI. CONCLUSION

Several researches only focused an energy as constrain, for security issues some standard algorithms are used to solve it by probability of node capturing. An African buffalo optimization based key management is proposed in this research to maintain the security level of all nodes by keeping a cluster heads. The ABO-KM provides effective selection and capturing the nodes activity. The dynamic changing of buffalo's location directly links the nodes and their respective updating mechanism. The risk factor is reduced and the security level of data transmission is improved. However, still there is a need to improve network applicability by several factors such as key management, applying the proposed scheme in key management scheme, implementing it in secure routing that focus on multi path routing. In addition to that the security solutions are focused based on the management overhead.

REFERENCES

[1]. Celozzi, C., Gandino, F., & Rebaudengo, M. (2013, June). Improving Key Negotiation in Transitory Master Key Schemes for Wireless Sensor Networks. In International Conference on Sensor Systems and Software (pp. 1-16). Springer International Publishing.

[2]. Chen, L., & Chen, L. (2014). An improved secure routing protocol based on clustering for wireless sensor networks. In Mechatronics and Automatic Control Systems (pp. 995-1001). Springer International Publishing.

[3]. Du, W., Deng, J., Han, Y. S., Chen, S., & Varshney, P. K. (2004, March). A key management scheme for wireless sensor networks using deployment knowledge. In INFOCOM 2004. Twenty-third AnnualJoint conference of the IEEE computer and communications societies (Vol. 1). IEEE.

[4]. Hussein, R. M., Ahmed, H. S., & El-Wahed, W. F. A. (2010, March). New encryption schema based on swarm intelligence chaotic map. In Informatics and Systems (INFOS), 2010 The 7th International Conference on (pp. 1-7). IEEE.

[5]. Khan, S., Ali, A., & Durrani, M. Y. (2013). Ant-Crypto, a Cryptographer for Data Encryption Standard. IJCSI International Journal of Computer Science Issues, 10(1).

[6]. Latiff, N. A., Malik, N. N. A., & Idoumghar, L. (2016, August). Hybrid Backtracking Search Optimization Algorithm and K-Means for Clustering in Wireless Sensor Networks. In Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C (pp. 558-564). IEEE.

[7]. Lee, Y. H., Phadke, V., Deshmukh, A., & Lee, J. W. (2004, August). Key management in wireless sensor networks. In European Workshop on Security in Ad-hoc and Sensor Networks (pp. 190-204). Springer Berlin Heidelberg.

[8]. Odili, J. B., Nizam, M., & Kahar, M. (2016). African Buffalo Optimization. International Journal of Software Engineering & Computer Sciences (IJSECS), 2, 28-50.

[9]. Rahman, M., & Sampalli, S. (2015). An efficient pairwise and group key management protocol for wireless sensor network. Wireless Personal Communications, 84(3), 2035-2053.

[10]. Rescorla, E. (1999). Diffie-Hellman key agreement method.

[11]. Sarkar, A., & Mandal, J. (2012). Swarm intelligence based faster public-key cryptography in

wireless communication (SIFPKC). Int. J. Comput. Sci. Eng. Technol.(IJCSET), 7, 267-273.

[12]. Sreelaja, N. K., & Pai, G. V. (2008, January). Swarm intelligence based key generation for text encryption in cellular networks. In Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference on (pp. 622-629). IEEE.

[13]. Sreelaja, N. K., & Vijayalakshmi Pai, G. A. (2011). Swarm intelligence based key generation for stream cipher. Security and Communication Networks, 4(2), 181-194.

[14]. Stinson, D. R. (2005). Cryptography: theory and practice. CRC press.

[15]. Wan, C., Zhang, J., & Huang, J. (2015). PKDIP: Efficient Public-Key-Based Data Integrity Protection for Wireless Image Sensors. Journal of Sensors, 2015.

Zhu, S., Setia, S., & Jajodia, S. (2006). LEAP+: Efficient security mechanisms for large-scale distributed sensor networks. ACM Transactions on Sensor Networks (TOSN), 2(4), 500-528

Analysis and Design of Pretend Based Security Campus Grid Computing Model for Arbaminch University

Daniel Tadesse[1] Dr J.R.Arun Kumar[2]
[1,] M.Sc Research Scholar, [2] Assistant Professor (advisor)
[1,2]Department of Computer Science and IT,
Arbaminch Institute of Technology, Arbaminch University, Arbaminch, Ethiopia.
[2]arunnote@yahoo.com, [1]daniel.ezanadan@gmail.com

***Abstract* —** A **Computational Grid is a set of heterogeneous computers and resources unfold across a couple of administrative domains with the purpose of presenting users uniform get access to these resources. A complete set of Grid utilization scenarios are presented and analyzed with reference to security requirements consisting of authentication, authorization, integrity, and confidentiality. With its application popularization, however, researchers and IT organization associations begin listening to the possible security issues from grid environments. Gird environments are based totally at the internet, they consequently face protection threats from numerous aspects, consisting of outside intrusions, inner assaults and pass-area name security problems, and many others. Kerberos integration protection and authentication system have been used to protect the resources of Grid computing. Even though the Kerberos V5 authentication system has a number of limitations and weakness. The main focuses on this research identify limitation and weakness of the Kerberos and redesign a well-defined multilevel security model for Arba Minch University Grid. This proposed design is called AMU classy security model and it enhances group policy and components to access the resources and data of grid network users with existing Kerberos authentication mechanisms. It is designed and validated by using Globus tool kit Grid security infrastructure (GSI). This research focuses on the integration of Kerberos GSI which complements the discussed above security vulnerabilities in Arbaminch University.**

*Keyword**: - Grid Computing, Kerberos, GSI, Malware, Authentication, AMU classy model, Globus tool Kit***

I. INTRODUCTION

State of the art and emerging scientific applications requires high performance computing [1]. It is not possible to solve large quantities of data and problems using a single high performance computer or a single computer cluster. Therefore it is required to connect distributed, heterogeneous high performance computers [6], computer cluster with high speed interconnection networks and integrate them into a high performance computing environment. This environment is known as Grid Computing. Large scale distributed environments couple computers storage system and other devices to enable advance applications such as distributed supercomputing, computer enhancement instruments and distributed data mining. Grid Computing Systems results from development of application for high performance computing. It includes dynamic resource requirements, use of resources from multiple administrative domains, complex communication structures and high performance requirements. Grid Computing Systems now attracts universities, institutions, governments, military and big companies. The characteristics of computational Grids leads to security problems that are not address by existing security technologies for distributed systems. The dynamic nature of the Grid can make it impossible to establish trust relationships between sites prior to application execution. Security is a latest topic today for the smart grid, and progresses are being done in this field every day. Most communications uses standard cryptographic algorithms AES-128 to protect the data on the network. Grid computing is a technique which provides high-performance computing; in this resources are shared in order to improve the performance of the system at a lower price. According to literature, “Grid computing is a system where multiple applications can integrate and use their resource efficiently” .It has three important categories: coordination of resources not under centralized control, use standard general purpose interface, and it delivers nontrivial quality of service.

II. LITERATURE REVIEW

In this literature review, major works about Grid computing and Grid security models that are highly related to this work are discussed & presents and analyzes the Grid security infrastructure that is implemented in various Grid environments. Then it presents a research project implemented using one of the Grid environments that uses present security

infrastructure. This research helped in better understanding of the Grid environment and security related issues. The section focuses on Review of existing Grid Security Structure. Security is important to all computer systems to enable administration and policy enforcement. These policies control the users of the system by specifying which user can access the system, what operation is allowed by each user, and protects the system of being compromised or misused. This is mapped to the basic security requirements of any system which are authentication, authorization (access control), integrity, privacy, and non-repudiation [11]. Implementing security mechanisms in the Grid is important to protect the large number of resources and users. The problem in implementing such security mechanisms in the Grid is that grid application may require access to multiple computational and data sources that may be geographically distributed and administrated locally and independently. Grid applications may involve hundreds of processes running on different resources and need to authenticate and communicate with each other securely.

These processes and resources may also join or leave dynamically which makes it impossible to initialize the security relationship between them at the application startup. All these problems complicate the implementation of a Grid security system and add new security problems not addresses by existing distributed security mechanisms such as Kerberos and the secure shell [12].

Related Works

Recently, lot of work has appeared in the literature on the problems of the computational grid. Varieties of problems have been studied and mentioned here. Krauter and Maheswaran (1997) proposed a grid architecture that is motivated by the large-scale routing principles in the Internet to provide an extensible, high-performance, scalable, and secure grid. Central to the proposed architecture is middleware called the grid operating system (GridOS). This paper describes the components of the GridOS. The GridOS includes several novel ideas (i) a flexible naming scheme called “Gridspaces", (ii) a service mobility protocol, and (iii) a highly decentralized grid scheduling mechanism called the router-allocator. Demchenko et al. [15] have commented that the Open Grid Forum (OGF) had been concentrating on short term security goals of achieving interoperability between presently deployed grid systems. They have further indicated that the main focus of the OGF had been on the primary security services and mechanisms such as authentication, authorization and web service protocol security. They have also shown that there is a gap between the Open Grid Security Architecture (OGSA) security model and services definition of the OGF and the practical grid implementations such as LCG/EGEE (LHC Computing Grid /Enabling Grids for E-Science), OSG (Open Science Grid) etc. The main reason for these gaps has been identified as the use of different grid middleware implementations. Hoeft and Epting [04] have discussed the Integrated Site Security for Grid (ISSeG) project carried out by the European Union in detail. ISSeG has been carried out as a part of the EU Framework Programme 6 by CERN of Switzerland, Forschungszentrum Karlsruhe GmbH (FZK) of Germany and STFC of the United Kingdom with specific responsibilities assigned to each partner. The ISSeG project aims to fulfil two main objectives. They are namely; to retain the site security for an effective working environment, while maintaining sufficient openness for scientific research and to ensure the three pillars of security confidentiality, integrity and availability of research and personal data. ISSeG proposes to have centralized resource management for faster detection and management of security breaches, integrated identity and resource management, and enhanced network connectivity management. It will develop and deploy security mechanisms and tools to implement effective security training, best practices and administrative procedures among all the stakeholders of the project. It can be seen that the ISSeG project focuses solely on centralized resource and event management. Though, centralized resource and event management has its own advantages in terms of efficient utilization of resources, performance will suffer when the system grows large and also limits the independence of partners. Also, ISSeG require all the sites and VOs to be homogeneous severely restricting the innovation within systems. Qiang and Konstantinov [14] have described a single sign-on infrastructure developed as a part of the NorduGrid Advanced Resource Connector (ARC) Grid middleware. They propose that the single sign-on with identity federation that can facilitate cross domain access would be more suitable for non-IT users than Public Key Infrastructures (PKI) to support mutual authentication using X.509 certificates. The proposed infrastructure totally eliminates the use of X.509 credentials and the users will need only their username/password combination to access any resource within the grid system. It also proposes to use Short Lived Credential Services (SLCS) for accessing grid systems that require X.509 credentials to access resources within their domains.

The temporary certificates issued by SLCS have a life time of 12 hours and hence it is proposed to use only local file permissions to protect the private keys instead of encrypting those using passphrases. This is one of the main shortcomings of this proposed mechanism as it can be exploited by IT users to attack the entire system including the ones that are protected using PKI. Li et al. (2006) have proposed an alternative security architecture that can be implemented in place of OGSA. They have also proposed that the grid security system is composed of two main components, namely; security rule definition and security rule implementation. The proposed framework can be extended easily compared to the GSI as it is developed independent of the grid system and loosely coupled to it. However the proposed mechanism suffers from the shortcoming that it stores the security data outside the grid system and hence can be attacked by malicious users easily.

III. PROBLEM STATEMENT

Security and privacy are two main concerns now days in the world. When we discuss about Grid Computing Infrastructure; immediately a question arises in our mind. Is that our data or application running on shared resources are secured enough and privacy has not been compromised? Grid computing systems environments require an ability to provide a secured running/ execution environment for applications where many users are working on different domains on the same platform. The grid should allow arbitrary code from untrusted users to legitimately share resources, while providing an active enforcement of the security and privacy policy of the shared resources. Kerberos optionally provides integrity and confidentiality for data sent between the client and server. Kerberos is not effective against password guessing attacks; if a user chooses a poor password, then an attacker guessing that password can impersonate the user. Similarly, Kerberos requires a trusted path through which passwords are entered. Currently, the Grid Security Infrastructure (GSI) does not have any intrusion detection system that can avoid any of the previously presented security leaks.

IV. RESEARCH METHODOLOGY

The sampling method was random judgment sampling. Twenty computers were randomly chosen from AMU computer facilities. The Desk tops chosen included Main Camps, Nechsare College of Medicine and Abaya Campus. Then questionnaires were administered to staff of the randomly selected staffs. The questionnaire was distributed to management (academic and administrative) and information technology staff of the universities. These categories of staff are likely to be responsible for taking decisions on the adoption of major technology facilities. The data gathering instrument used in the research is direct observation and questioners that are fully answered but it's not satisfactory for our research we mainly used our simulation tools results from the system that has been configured earlier. A total of fifty questionnaires were distributed to the fifteen AMU staffs for sample for this study. Forty-four of them were properly filled and used for analysis purposes. This represents 95% response rate. The questionnaire consists of two sections. The first part summarizes demographic information including information on if they heard about Grid computing and availability of grid computing supporting infrastructure, and implementation of full grid computing. In the direct observation methods for collecting data regarding current status of AMU Grid computing and Grid computing secured resource allocation test has been done in AMU ICT on installed server and on Desktop that are connected with the Grid computing. We have collected data related to the resource sharing in a secured manner and testing it on the simulation of the grid will be presented in the next chapter of this research and in this process several evidences are gathered regarding the Grid security.

This research is implemented by the Globus Toolkit (GT) [4] is an open community-based, open source set of services and software libraries that was originally developed at a national laboratory. Globus provides standard building blocks and tools for use by application developers and system integrators. It has already gone through four versions in the last few years: the original version in the late 1990s, GT2 in 2000, GT3 in 2003, and GT4 in 2005. GT2 was the basis for many Grid developments worldwide. GT3 was the first full-scale implementation of Grid infrastructure built upon Web Services by way of the GGF's OGSI intermediate layer [5]. GT4 is the first implementation that is compliant with mainstream Web Services as well as Grid services based on WSDL [6] and WSRF [7].

V. PROPOSED AMU-CLASSY GRID SECURITY MODEL

An AMU-CLASSY GRID is aware of agreements with other Campus Grids acts as a Grid selector by selecting a suitable Grid able to provide the required resources; and replies to requests from other Grids, considering its policies. IGGs with pluggable policies enable resource allocation across multiple Grids. An AMU-Classy Grid is chosen and maintained by a Grid based on internal criteria. It is also interacts with other entities including Grid Information Services (GISs), resource discovery networks, accounting systems and resource managers within peered Grids. A GIS provides details about the available resources; and accounting systems provide information on shares consumed by peering Grids. It is illustrated how our new model works in explained in figure

When a user wants to gain access to a server, the server needs to verify the user's identity. Consider a situation in which the user claims to be, for example, Daniel.tadesse@amu.edu.et. Because access to resources are based on identity and associated permissions, the server must be sure the user really has the identity it claims. When the user logins to his or her machine. The principal is sent to KDC server for login, and the KDC server will provide TGT in return. Next KDC server searches the principal name that means *Daniel.tadesse@amu.edu.et* in the database, on finding the principal, a TGT is generated by the KDC, which will be encrypted by the users key, and send back to the user. When the user gets the TGT, the user decrypts the TGT with the help of KINIT (with help of the users' key). An important fact to note here is that, the client machine stores its key on its own machine only and this is never transmitted over wire.

The TGT received by the client from the KDC server will be stored in the cache for use for the session duration. There will always be an expiration time set on the TGT offered by the KDC server, so that an expired TGT can never be used by an attacker.

Finally the client has got TGT in hand. If suppose the client needs to communicate with some service on that network, the client will ask the KDC server, for a ticket for that specific service with the help of TGT. Even if the Kerberos protocol authenticates the user's identity, it does not authorize access. This is an important distinction. Tickets in other contexts, such as drivers' licenses, often both prove identity and authorize actions or access. A Kerberos ticket only proves that the user is who the user claims to be. After the user's identity is verified, the Local Security Authority will authorize or deny resource access. Kerberos messages are encrypted with various encryption keys to ensure that no one can tamper with the client's ticket or with other data in a Kerberos message. As we have explained in short in the above paragraph how Kerberos authentication is working we will now give some clue on how to integrate it with GSI. The next part will explain it step by step.

The GSI focuses on two main issues. The first one is to provide an authentication mechanism between users, user processes, and resources. This authentication will enable different local security policies to be integrated into a single global framework. The second issue is to enable the application of local access control mechanisms. The Model consists of different Policies, like the Smart Grid Certificate (SGC) and the Grid Policy Manager (GPM). The SGC is a process that acts on behalf of the user for a limited time to manage a computation session. The purpose of the SGC is to enable single sign-on. The SGC is created first through user authentication using Kerberos.

The user first gains access to the machine on which the SGC will be created AMU-SGC (CA). Then the user uses his credential (**CA**) to create temporary credential (SGC-**CA**). Finally the CA process is created and given its temporary credentials (SGC-**CA**) that will be used for further authentications and acting on behalf of the user. The temporary credentials consist of a tuple – containing various information such as the valid interval of this credential, authorized actions, user ID, and so on – signed by the user credentials. It was possible to give the user credentials to the SGC-**CA** to enable single sign-on which is very simple solution. Alternatively, temporary credentials were used for two reasons: To protect the user credentials. Because giving the user credentials to the SGC-**CA** and then to processes acting on behalf of the user increases the probability of user credentials being hacked and misused. To control the SGC-**CA** and the user processes by delegating a subset of the user rights – the signed tuple – to these processes through the temporary credentials and so reducing security risks. After the SGC-**CA** is created, the user can leave the computation.

Interoperability with local security solutions is achieved through the Smart Resource Allocation (SRA) that acts as an agent translating between inter-domain security operations and local intra-domain mechanisms. The SGC-**CA** contacts the Grid Policy Manager (GPM) and Smart resource allocator (SRA) to allocate the resource. First the SGC-**CA** and the SGC authenticate with each other (mutual authentication). Then the SGC-**CA** sends assigned request to the SRA. the SRA checks if this user is allowed – according to local security mechanisms – to access the resource. If the user is authorized the user and resource proxies negotiate together to create the process temporary credential (**CA**). Finally the SRA allocates the resource and passes (**CP**) to the newly

created processes. The process credential (CP) facilitates communication by enabling the authentication between the SGC-CA and the processes and between the processes themselves if they exist in different domains. It also enables a process to acquire more resources as shown Fig 1. In the below
in which it first mutually authenticate with the SGC-CA. Then it passes a signed request to the SGC-CA that
.

the secure group – after accepting the request – allocates the resource through SRA

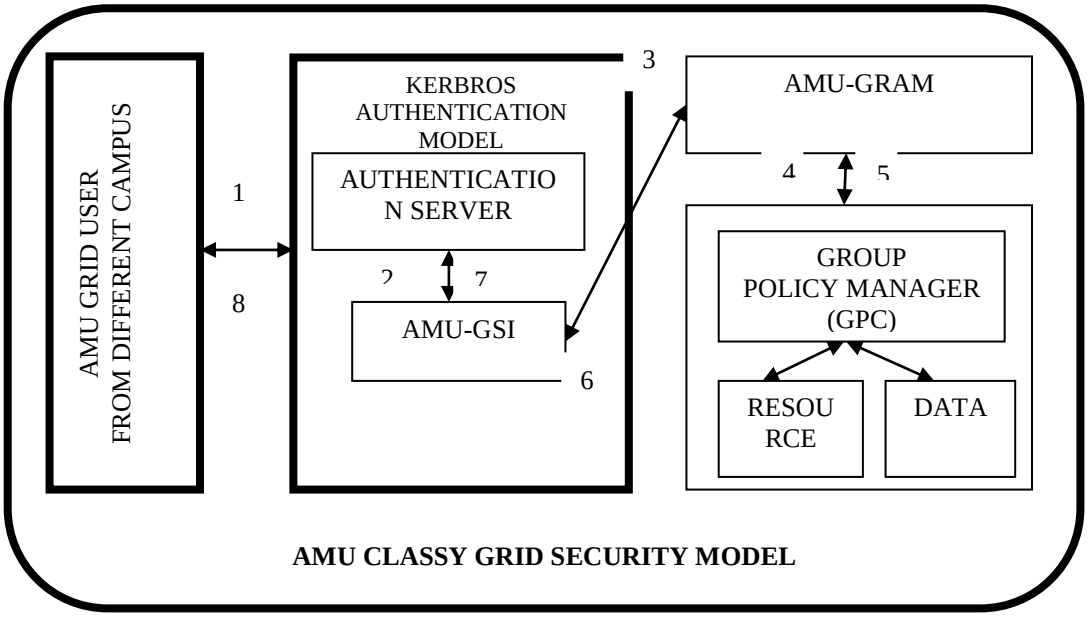


Fig 1. Proposed AMU classy grid security model

VI. EXPERIMENTAL RESULTS

The email that is register as *root@globus.amu.edu.et:daniel.ezanadan@gmail.com* gets the CA certificate and the certificate has an expiration date. Keep in mind that once the CA certificate has expired, all the certificates signed by that CA become invalid. A CA should regenerate the CA certificate and start re-issuing CA –setup packages before the actual CA certificate expires. This can be done by re-running this setup script. Enter the number of DAYS the CA Certificate should last before it expires Image Result from the above source code User Certificate Server side CA for user *Daniel.ezanadan@gmail.com*

```
   C e r t i f i c a t e   A u t h o r i t y   S e t u p

This script will setup a Certificate Authority for signing Globus
users certificates.  It will also generate a simple CA package
that can be distributed to the users of the CA.

The CA information about the certificates it distributes will
be kept in:

/var/lib/globus/simple_ca

It looks like a CA has already been setup at this location.
Do you want to overwrite this CA? (y/n) [n]: y

The unique subject name for this CA is:

cn=Globus Simple CA, ou=simpleCA-globus.amu.edu.et, ou=GlobusTest, o=Grid

Do you want to keep this as the CA subject (y/n) [y]: y

Enter the email of the CA (this is the email where certificate
requests will be sent to be signed by the CA) [root@globus.amu.edu.et]: daniel.ezanadan@gmail.com
    The CA certificate has an expiration date. Keep in mind that
            once the CA certificate has expired, all the certificates
            signed by that CA become invalid.  A CA should regenerate
            the CA certificate and start re-issuing ca-setup packages
            before the actual CA certificate expires.  This can be done
            by re-running this setup script.  Enter the number of DAYS
            the CA certificate should last before it expires.
[default: 5 years 1825 days]:
```

Fig.2. Experimental Evaluation of Proposed AMU CA using Globus Tool Kit

The above Fig 2. contains the globus certificate (ca) with the ca owner of "simpleca-globus.amu.edu.et" the ca information about certificate it will be distributed is kept inside the "var/lib/globus/simple_ca path directory and related policy files needed to allow the globus toolkit to trust the ca with the subject name: "${_casubject}" as well as request certificates from that ca.in the above figure 7 the email address of the ca that means the email where certificate requests will be sent to be signed by ca *root@globus.amu.edu.et|daniel.ezanadan@gmail.com* after receiving the ca the user or the client uses that ca to communicate with the grid. Amu smart grid resource allocator manger (amugram) will check and forward it to the resource information service (broker).the resource information service (broker) will assign the user to the exact point that means, if the user requests to access the hard ware resource the broker will send it hardware resource and if the user want to access (use) software it will assign to the software resource and the vices versa

VII. CONCLUSION

This thesis has investigated some critical aspects in research of security issues in grid computing. Grid computing presents a number of security challenges that are met by the globus tool kit's grid security infrastructure. Several grid architectures have been proposed in last ten years. Globus toolkit implements the emerging open grid services architecture with amu glassy security model. It's gsi and classy model implementation takes advantages of this advancement to enhance on the security model used in earlier version of the ca's. The security in grid environment is achieved through the implementation of the various security measures such as authentication, authorization, and data privacy. This research study focused an overview of security issues concerned mainly in certificate authority. In addition, in some specific issues, new technologies can be secured than older ones due to that the design of the new solutions can be more suitable to avoid the security problems and will make this task easier. Further strong authentication certificate authority must be developed in future for the improvement of security resources in the grid computing

REFERENCES

[1] Ian Foster, Carl Kesselman (eds). (1 November 1998).The Grid: Blueprint for a New Computing Infrastructure (1st edition). San Francisco, USA: Morgan Kaufmann publishers.

[2] Foster, I., Kesselman, C., Tuecke, S.. (2001). The Anatomy of the Grid: Enabling Scalable Virtual organizations. International Journal of Supercomputer Applications, 15(3).

[3] Humphrey, M., Thompson, M., Jackson, K. (March) 2005.Security for grids. Page 644-652 Proc. Of IEEE

[4] Analysis on Grid Security Patterns Based on PKI Zhengli Zhai, Yan Qiao, Mingwen Shao

[5] I Foster, C Kesselman, G Tsudik, and S Tuecke, A Security Architecture for Computational Grids, in Proc 5th ACM Conference on Computer and Communications Security. 1998. p. 83-92.

[6] Howard Chivers, John A. Clark, and Susan Stepney, Smart Devices and Software Agents; the Basic of Good Behaviour, in Proceedings of the first International Conference on Security in Pervasive Computing. 2003, Springer-Verlag: Boppard, Germany.

[7] R. Lock, Prof I. Sommerville. Grid security and its use of X.509 certificates. Lancaster University.

[8] Cryptography and network security principles and practice fifth edition William Stallings

[9] Virtual Organizations. International Journal of High Performance Computing Applications, 15 (3). 200-222. 2001. **www.globus.org/research/papers/anatomy.pdf**.

[10] M. Baker, R. Buyya, and D. Laforenza, The Grid: International Efforts in Global Computing, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR 2000), l'Aquila, Rome, Italy, July 2000

[11] Jianmin Zhu and Dr. Bhavani Thuraisingham, "Secure Grid Computing", IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.8B, August 2006, pp. 216-229.

[12] R. Butler, D. Engert, I. Foster, C. Kesselman, S. Tuecke, J. Volmer, and V. Welch, A National-Scale Authentication Infrastructure. IEEE Computer, Vol 33, No 12, pp 60-66, 2000.

[13] K. Nichols, V. Jacobson, and L. Zhang (1999), "A Two-bit Differentiated Services Architecture for the Internet," RFC 2638, July 1999.

[14] K. Kilkki (1999) Differentiated Services for the Internet, Macmillan Technical Publishing. The Globus toolkit, http://www.globus.org/toolkit/about.html.

[15] Watkins, R. 2015. IBM Grid Computing & Virtualization, GRID @ Asia Conference, Seoul, Korea, December11-13

Towards Integrating Data Mining and Knowledge Based System: The Case of Network Intrusion Detection

Abdulkerim M. Yibre[1], Million Meshesha[2]

[1] Department of Computer Engineering Selçuk University Konya, Turkey, abdukerimm@selcuk.edu.tr
[2] School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia,

Email: million .meshesha@aau.edu.et

***Abstract* — In this study, rule based intrusion detection and advising knowledge based system (RIDA-KBS) is proposed. The system is designed aiming at utilizing hidden knowledge extracted by employing induction algorithm of data mining specifically JRip from sampled KDDcup'99 intrusion data set. The integrator application is developed to link the model created by JRip to knowledge based system so as to add knowledge automatically. The integrator reverses the IF …THN rule syntax of JRIP to PROLOG understandable format. The performance of the system is evaluated by preparing test cases. Twenty test cases are prepared and provided to domain experts. The same instances are tested on RIDA-KBS and has scored 80% overall performance which is a promising result. But further exploration needs to be done to refine the knowledge base and boost the advantages of integrating data mining induced knowledge with knowledge based system.**

Key words- **intrusion detection, knowledge based system, data mining, and integration**

I. INTRODUCTION

These days the world of computing has encountered with the ever-increasing likelihood of unexpected downtime due to various attacks and security breaches (Ghorbani, 2010). Network downtime effects in financial losses and harms the credibility of organizations. Minimizing or eliminating the unplanned and unexpected downtime of networks can be achieved by identifying, prioritizing and defending against misuse, attacks and vulnerabilities.

Intrusion is a type of attack on a computer system that attempts to bypass its security mechanism. Attacks can be done by an outsider who attempts to access the system, or an insider who attempts to gain and misuse un-authorized privileges (Srinivasulu, Nagaraju, Kumar, & Rao, 2009).

Intrusion detection is the process of identifying and responding to malicious activities targeted at computing and network resources (Ghorbani, 2010). Intrusion detection systems are core elements in network security infrastructure. They examine system or network activities to find potential intrusions or attacks and trigger security alerts for the malicious actions.

In the struggle of developing effective Intrusion detection, data mining techniques has been used by the computing world (Barbara, Couto, Jajodia, Popyack, & Wu, 2001; Dokas et al., 2002; Ektefa, Memar, Sidi, & Affendey, 2010; Lee, Stolfo, & Mok, 1999; Srinivasulu et al., 2009; Tiwari, Tiwari, & Yadav, 2013). However, (Domingos, 2007) stated that, there is a gap between the results a data mining system can provide and taking action based on them. In addition, commercial network intrusion detection system mostly generates alarms when they get attacks according to their knowledge base and the action to be taken is left to the network administrator. This study is not restricted to merely extracting hidden knowledge from KDDcup'99 intrusion dataset; rather it adds value to the extracted knowledge by integrating it with knowledge based system.

Developing knowledge base system is paramount to identify different types of attack and give advice accordingly to help the administrators which action to take. The integration of data mining induced knowledge with knowledge based system allows utilizing interesting and previously unseen knowledge extracted from data mining models for knowledge base system. This again lessens the problems of commercial intrusion detection systems from merely notifying while detecting attacks by providing advice and information about the detected network attacks.

Knowledge acquisition is the one of the fundamental components in knowledge based system. Knowledge can be acquired manually where knowledge is collected from some knowledge expert. The other mechanism is automatic knowledge acquisition where knowledge is acquired using computers through extracting from data we have at hand. In this respect, data mining techniques became the most used in the recent years (Oprea, 2006; Turban, Aronson, & Liang, 2005). Data mining has been proposed for extracting hidden and previously unknown knowledge from datasets by different researchers (Kamber, 2006).

The objective of this study is integrating knowledge acquired using data mining with knowledge based system through designing prototype rule based intrusion detection and advising knowledge based system. In order to achieve that, first we prepared sample of intrusion data set from KDDcup'99. Then used number of rule-based classification algorithms. The algorithms are compared according to their classification accuracy in identifying an attack example from normal ones. After that knowledge was acquired using the selected algorithm. We developed an integrator for automatically building knowledge base gained from classification algorithm. And finally, constructed prototype knowledge based system.

The paper was organized as follows. The second section was about the materials and methods, third section discusses the classification algorithms, the fourth one is about data mining techniques for intrusion detection. The fifth section was about knowledge acquisition, the sixth section was about integration, results and discussions and the final section was about conclusion and recommendation.

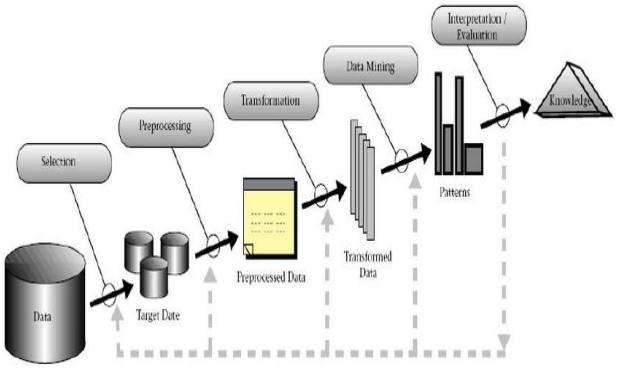


Figure 1 KDD process components (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

II. MATERIALS AND METHODS

Knowledge Discovery in Database (KDD) model was used for the data mining task. This procedure was used for extracting and refining valuable knowledge from large databases [5]. KDD has been used by different researchers to discover knowledge from large collection of records. It has seven steps in which data mining is at the fifth step.

Data source: KDDcup'99 was our data source. KDDcup'99 is a data set which has been in use since 1999 for evaluation of anomaly detection methods. The dataset was built based on the data captured in DARPA'98 intrusion detection system evaluation program. KDD training data set consists of 41 features and is labeled normal, probe, DOS, R2L and U2R (Tavallaee, Bagheri, Lu, & Ghorbani, 2009).

Attack type	Number of Instances	Percentage
Normal	22,000	61.40%
Probe	5775	16.10%
DOS	7892	22.05%
R2L	102	0.20%
U2R	9	0.00%
Total	35,788	100%

Table 1 Types of network attacks

Before removing duplicates		After removing duplicates	
Normal	595,797	Normal	559,276
Attacks	452,778	Attacks	55,171
Total	1,048,575		614,447

Table 2 Number of 'normal' and 'attacks'

Data preprocessing:- According to (Kamber, 2006), real world databases are prone to noisy, missing and inconsistent. This is due to their huge size and their likely origin from multiple sources. Low quality data will lead to low-quality mining results. The KDDcup'99 dataset has inherent problems in which the most important one is the existence of redundant instances (Tavallaee et al., 2009). From the collected 1,048,575 instances of KDD dataset for this study, 58.5% of them are found redundant. Therefore, before the actual mining task is performed, these instances are removed at the data preprocessing stage. The number of instances with their class name is depicted under Table 1.

KDD dataset is very huge in size; even after removal of redundant instances. The remaining dataset is so huge which requires time and memory space during the mining process. Hence, the sampling is found to be paramount. To generate or extract knowledge from the dataset and considerable samples were taken. While sampling, all instances of R2L and U2R are *purposively* taken since their size is so small compared to others. The number of instances included in the sample for *normal*, *probe* and *DOS* is based on their proportion on the cleaned dataset.

Knowledge representation- Rule based knowledge representation approach was used to represent knowledge. A rule is a conditional statement that links given conditions to action or outcomes

(Abraham, 2005). Rules are built in the form of *if-then* format. These *if-then* rules statements are used to formulate the conditional statements that constitute knowledge base. It is very easy to read, easy to interpret and easy to generate. In addition, it also classifies new instances rapidly (Datta & Saha, 2011).

We used WEKA as a mining tool. It is proven to be powerful for data mining. It contains tools for data preprocessing, clustering, regression, classification, association rules and visualization. WEKA wa written in Java language and contains Graphical User Interface for interacting with data files and producing visual results. It can be embedded like a library in application since it has Application Page Interface.

In addition, PROLOG was used to represent rules in the knowledge base and to construct the prototype system. PROLOG is an AI programming language belonging to the group of logic programming. It is a declarative language in a sense that computations are carried over by running queries over the relations defined as rules and facts (Shapiro, 1986). PROLOG has data type called *term*, which can be an atom, number, variable or a compound term.

III. RULE BASED CLASSIFICATION

Classification is a process of building model that designate data class and used to predict the class of objects whose class label is not known. That means classification is a two-step process. Firstly, a model is built describing a labeled set of data classes. This is the learning step. Secondly, the model is used to predict unlabeled data set. Analysis of set of data is the base for the model. The data is grouped in to two parts: one for building the model; the other for testing the model.

Rule base classifiers group instances by using a set of "IF…THEN" rules.

Rule: IF (condition) THEN (X)

Where,

- Condition is a conjunction of attributes.
- X is a class label.

In rules, there are Left Hand Side (LFS) also called condition and Right Hand Side (RHS) also called the conclusion. A given rule *R* covers an instance *I* if the attributes of the instances satisfy the condition of the rule (Datta & Saha, 2011). Rule base classification has been in use for network intrusion detection (Liao, Lin, Lin, & Tung, 2013; Mitchell & Chen, 2013; Modi et al., 2013; Tiwari et al., 2013). We found rule based classifiers appropriate for classification due to the fact that rules are mandatory to construct knowledge based system. JRIP, was the selected classification algorithm.

JRIP is a prepositional rule learner, i.e. Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Rules in this algorithm are generated for every class in the training set and are then pruned. The discovered knowledge is represented in the form of IF-THEN prediction rules.

JRIP forms two subgroups namely a growing set and a pruning set while constructing rules. The rule is made up from instances in the growing set. Initially the rule set is empty and the rules are added incrementally to the rule set until no negative instances are covered. Next to this, in order to increase the accuracy of rules the algorithms substitutes or revises individual rules by using reduced error pruning. To prune a rule the algorithm takes in account only a final sequence of conditions from the rule and sorts the deletion that maximizes the function (Chauhan, Kumar, Pundir, & Pilli, 2013).

IV. DATA MINING TECHNIQUES FOR INTRUSION DETECTION

Data mining empowers to generate interesting knowledge or patterns in a huge collection of data set (Oprea, 2006). It is the most important place out of the seven steps of the knowledge discovery process (see Fig. 1). It is composed of analytical technique drawn from a range of disciplines such as numerical analysis, pattern matching, and areas of artificial intelligence; including neural network, machine learning and genetic algorithm (Jackson, 2002). The availability of enormous amount of data and an increasing need for changing it in to useful information and knowledge made data mining to get a great deal of attention. In recent years, the computing and network security community used data mining techniques for identifying network attacks (Chauhan et al., 2013; Dokas et al., 2002; Ektefa et al., 2010; Ilgun, Kemmerer, & Porras, 1995; Lee et al., 1999; Mitchell & Chen, 2013; Modi et al., 2013; Srinivasulu et al., 2009; Tiwari et al., 2013). In Data mining based intrusion detection systems there are two groups namely; misuse detection and anomaly detection systems. In misuse detection a model is trained with pre-labeled data to differentiate normal and attacks, where classification algorithms are implemented for the task (Kamber, 2006). Misuse is successful in detecting an attack which is previously known. But it has a drawback in detecting new intrusion types. Contrary to this, the anomaly detection initially launches a model of normal system behaviors, and anomaly events or attacks are identified by comparing with the normal behavior.

It has capability to identify previously unknown attacks. However, the false alarm rate is high.

V. KNOWLEDGE ACQUISITION

Knowledge acquisition is core component is knowledge based system development. The critical component of Knowledge Based system is the knowledge base which contains rules and facts. Rule-based systems are mostly used in knowledge representation.

Classifier	Correctly classified instances		Incorrectly classified instances	
	No.	%	No.	%
PART	35,735	99.88	43	0.12
JRip	35,737	99.89	41	0.11
REPTree	35,646	99.63	132	0.37
J48	35,732	99.87	46	0.13

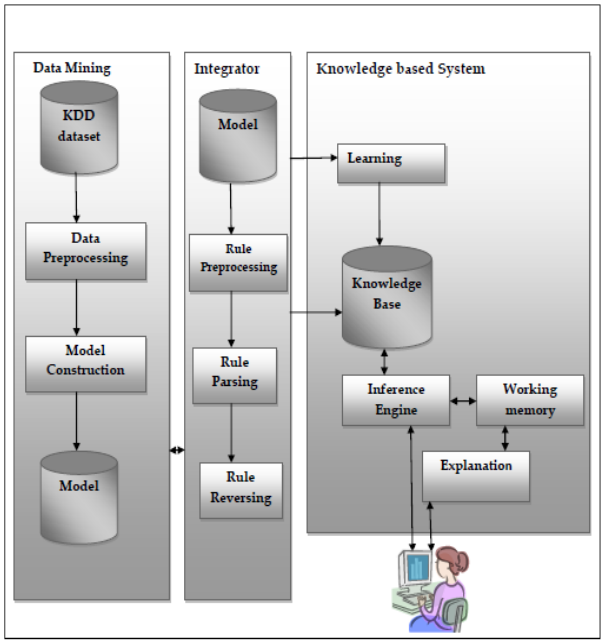
Table 3 performance of classifiers.

Classifier		class				
		Normal	Probe	DOS	U2R	R2L
PART	Precision	99.9	99.80	99.90	94.60	55.60
	Recall	100	99.80	100	85.30	55.60
	F-measure	99.90	99.80	99.90	89.70	55.60
JRip	Precision	99.90	99.80	99.90	98	75
	Recall	99.90	99.80	99.90	95.10	66.70
	F-measure	99.90	99.80	99.90	96.50	70.60
REPTree	Precision	99.90	99.70	99.10	89.60	75
	Recall	100	99.90	99.90	84.30	66.70
	F-measure	99.90	99.50	99.90	86.90	70.6
J48	Precision	99.90	99.90	99.90	93.90	33.30
	Recall	100	100	100	75.50	22.20
	F-measure	99.90	100	100	83.70	26.70

Table 4 precision recall and F-measure in percentage

Knowledge can also be acquired from large collection of dataset by using knowledge discovery tools. This type of knowledge is called hidden knowledge. For that, four rule generating algorithms namely; JRIP, REPTree, PART and J48 purposively selected. We used 35,788 instances for our experiment and 10-fold-cross validation test mode in WEKA was employed, where the total number of instances were divided into ten groups among them nine of them are used for training and one of it for testing.

The classification algorithm showed slight difference in terms of performance (shown under Table 3 and 4). Prediction accuracy shows us the general classification accuracy of the algorithms. Apart from prediction accuracy, classifiers are also evaluated to measure how they correctly classified each class to their correct class or incorrectly classified to another class. Hence, to evaluate the performance of the classifiers employed in this study True Positive rate, False Positive rate, Precision, Recall and F-measure are used. Due to its best result in accuracy, precision recall and F-measure and suitability of the rules generated, we selected JRIP algorithm.



VI. INTEGRATION OF DATA MINING RESULTS WITH KNOWLEDGE BASED SYSTEM

Following the knowledge acquisition, we automatically integrated knowledge base system. JRIP algorithm generates 23 rules which in the form of *IF(condition) THEN (conclusion)* format. For example:

(is_guest_login=1)and(duration<=1)=>class=R2L

Figure 2 General frame work of integration of data mining with knowledge based system

The condition part contains attribute, comparison operators equal to (=), less than or equal to (<=) or greater than or equal to (>=) and

value for the attribute. Two or more conditions are joined by 'and'. After the conditions the '=>' symbol(- which means implies) follows. The conclusion part of the rule one of the class names (normal, probe, DOS, U2R and R2L).

However, PROLOG is a backward chaining language. Backward chaining attempts to reach goals in the order in which they appear in the knowledge base (Krishnamoorthy & Rajeev, 1996; Shapiro, 1986). In PROLOG statements conclusion comes first, after that condition follows. Apart from that, the characters used for joining condition, terminating statement and the character used to mean implies are different (see table 4). This behavior made the integration process challenging, in that we could not directly send the rules to the knowledge base. Therefore, we come up with building an algorithm for making PROLOG suitable rules. The algorithms are for tokenize rules, parsing rules and facts and reversing the rules.

A. *Tokenize rules*

A given JRIP rule contains special characters, attribute names, comparison and logical operators. The tokenization process, as shown under Table 5, focuses on removing characters which are undesirable, replacing some special characters by PROLOG equivalent character.

Token	JRip tokens	PROLOG equivalent	Tokenization option
Special character	'(', left brace ')', right brace	Same	Replace by empty space
Comparison operator	>,<,=,>= <=	Same =<	Keep them Replace by '=<'
Logical operator	AND	,	Replace by ','

Table 5 Tokens in JRIP and PROLOG rules

The conjunction operator 'and' is replaced by its PROLOG equivalent ',' bearing same meaning and function i.e. joining two conditions. The '=>' is replaced by ':-'which means *IF* in PROLOG. In addition, the token 'class=' is replaced by 'attack' to make it predicate for head of rules. The algorithm is depicted in Fig 4.

B. *Parse rules and facts*

In this section, parsing is analyzing the components of JRIP rules. A given rule is composed of *condition* and *conclusion*. If the condition is evaluated true then the conclusion is executed. The *condition* part is also divided into one or more conditions. In case there are two or more conditions, they are connected by logical operator (AND). A condition is composed of an *attribute*, *comparison operator* and *value*. An *attribute* is a property or characteristic describing about something for example, duration, service, flag etc. are attributes describing a certain network incident. *Comparison operator* is used for comparing an attribute with value which can be number or string. Fig 5 show the algorithm of parsing rules

C. *Reverse rules*

The reverse rule stage is used to exchange the place of Left Hand Side (LHS) of the rule and Right Hand Side (RHS) of rule. The rationale here is to build rules in PROLOG understandable structure. Hence, the JRip rule order must be reversed (see Fig 6 for the algorithm).
For example; Given a set of attribute A_i for i=1 to m, and respective value V_i for attribute A_i, the original rule and reversed rule are shown below.

Rule: $(A_1=V_1)$ and $(A_2=v_2)$…, and $(A_m=V_m)$ THEN (conclusion)
Reversed rule: (*conclusion*):- $(A_1=V_1)$ and $(A_2=v_2)$…, and$(A_m=V_m)$.

D. Normalize rules

The normalization stage is to change all tokens which are already reversed into lower case. The class names from the original rule are in upper case. Since PROLOG understands a token which starts in or is totally in upper case as a variable, all the tokens are normalized to lower case.

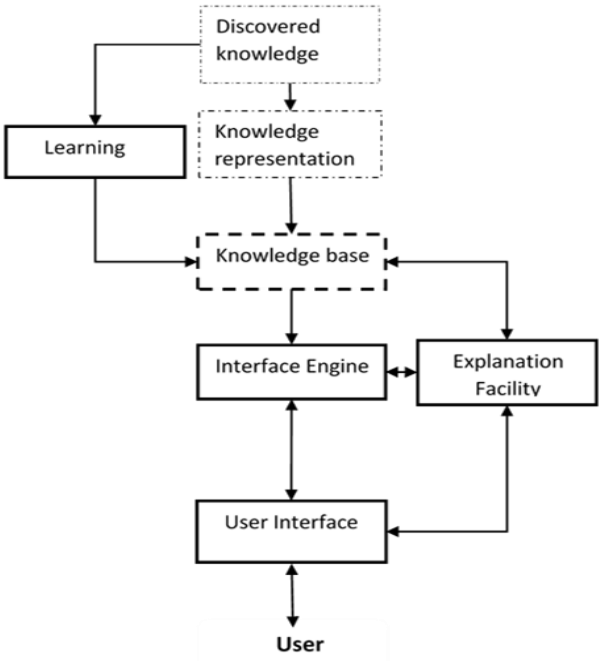


Figure 3 architecture of RIDA-KBS

```
function tokenizer(line)
        While ! endofline
         if length of line >0
                remove unwanted tokens
                 if token='and' then
                        replace it by comma(,)
                else if  token is '=>' then
                        replace it by ':-'
                else if token is 'class=' then
                        replace it by 'attack'
                else if token has
                   format '(digit/digit)' then
                remove it
        end if
end while
```

Figure 4 rule tokenizer algorithm

```
function rule_parser(line)
        read rule
        RULE_COMPONENT=split rule by '=>'
        CONDITION= left_of '=>'
        CONCLUSION=right_of '=>'
        CONDITIONS =CONDITION split by 'and'
        Loop:
            i=0 up to number_of(and)
            ANTECEDENT[i]=CONDTIONS[i]
            ATTRIBUTE [i]= left_of_comparision_operator
            VALUE [i]= right_of _comparion_operator
        end loop
```

Figure 5 parsing algorithm

```
Function rule_reversor(line)
 While end of line
       If token is the last token then
          If token_length not equal to zero then
           Reversed_rule=( token before the last_token AND
                           last_token  AND
                           third token from last_token )
              break
           end if
         end if
 end while
for i=0  till (LENGTH_OF_LINE) - 1
if token is not (LENGTH_OF_LINE) - 3
    if token is at first position
            concatenate (Reversed_rule
                    token at first position and opening brace
                    token next to first position token
                    token at third position
                    closing brace
                    space)
    else
            concatenate (Reversed_rule
                     comma(,)
                     token at first position and opening brace
                     token next to first position token
                     token at third position
                     closing brace
                    space)
             increment i by 4
    else
              concatenate (Reversed_rule, current_token
                 jump to next token
    end if
end if
end
```

Figure 6 rule reversing algorithm

VI. IMPLEMENTATION

So far, knowledge base was built automatically as rules and facts. The prototype Rule Based Intrusion Detection and advising Knowledge Based System (RIDA-KBS) is capable of diagnosing network incident as normal, DOS, U2R or R2L. The detection process is undertaken by interacting with the user by presenting a series of questions for the user. The system displays question containing attributes with their values of network incident. The user is expected to reply for the questions or ask explanation for the questions. Based on the question and answer the system identifies the incident as normal or an attack (see Fig.7). In addition, information related to each identified incident is displayed to assist the user in decision making. The user can either allow or deny a particular network incident. Discovered knowledge (accomplished at knowledge acquisition stage), knowledge representation and knowledge base are drawn as dashed line (in Fig. 3) to portray that they are previously undertaken. The learning component is included in the architecture. Learning is basically the capability of the knowledge based system to incorporate new rules or facts or both into its knowledge based system.

The rules change as the number of instances in the sampled data set changes. The designed knowledge base was able to accommodate these changes and use them in its diagnosis of network attacks. This accommodation of new rules by the KBS is called learning.

```
Is root_shell>=1 :?( what/yes/no) yes.
Is duration>=25: ?(what/yes/no) no.
Is num_failed_logins>=1 :?( what/yes/no) no.
Is service=ftp_data: ? (what/yes/no)  yes.
Is flag=sf:? (what/yes/no) yes.
The type of network attack is r2l (how)?
```

VII. RESULTS AND DISCUSSION

Under section five we presented the performance of JRIP algorithm. Those rules are automatically converted into rules and facts for the knowledge base. Here we can deduce that the performance of prototype system is also the same as JRIP rules.

However, In order to assure that the prototype RIDA-KBS meets the requirements for which it was developed for, it has to be tested with domain expert evaluation too. Therefore, system performance testing was done by preparing test cases. The test cases were samples of intrusion instances taken from KDDcup'99 intrusion data set. The instances include 20 attributes with their respective values. The test cases, which were unlabeled intrusion instances are delivered to domain experts to label them as normal, probe, U2R, R2L and DOS.
Considering the numbers of attributes and the time it consumes to label it manually, we prepared only 20 test cases/instances (7 normal and 4 probe, 4 DOS, 4 R2L and 1 U2R) for system performance testing. Attributes of instances with their respective values describe the behavior of certain network incident. Based on the attributes and their respective value, domain experts labeled the instances. Here our assumption was those instances labeled by domain experts as positive instance. The same set of test instances are provided to RIDA-KBS and the outputs are compared to the domain experts' judgment.

Confusion matrix is used for comparing the performance of RIDA-KBS with domain experts' judgment. In confusion matrix, the entries in the matrix indicate the number of attacks labeled as let's say; attack X by domain experts and detected as attack X or attack Y by RIDA-KBS. Accuracy, Precision, Recall, F-measure, True Positive rate and False Positive rate measures are used in performance testing.

Prediction accuracy measures the proportion of instances that are correctly classified by the classifier. In contrary to Predictive Accuracy TP rate and FP rate values do not depend on the relative size of positive and negative classes (Bramer, 2007). TP rate is the proportion of positive or correctly classified instances as positive or correct instances. FP rate also called False Alarm, measures the proportion of negative instances that are erroneously classified as positive. Precision measures the proportion of instances that are classified as positive that are really positive.

It is clear that each type of attacks has their own way of attacking and causing damages to the victim computer. Identification of each class of attacks to their correct class is important to provide proper advice to network administrators so that they can take appropriate measures. But as shown in the confusion matrix, 3 out of 4 normal instances are incorrectly classified as attacks. One instance out of 13 attack test cases/instances are incorrectly classified to another attack classes. We believe that the system's identification of this instance as an attack as strength though it is not to its correct class. The problem would have been if it were identified as normal instances.
In general, the system has correctly classified 16 test instances out of 20 to their correct class, which means 80% detection accuracy. But, this measure alone is not enough to measure performance of the knowledge base system, as this only tells us the overall performance. Hence Precision and Recall are also employed to evaluate system performance apart from detection accuracy.

So far promising result was achieved. However, the interaction of the users with the system was not easy due to the command line interface. Whenever there was a situation to rerun the system, it was highly likely that new rules were generated and the knowledge base was also updated accordingly. This made it difficult in keeping quality rules which were previously used.

		RIDA-KBS					
	Class	**Normal**	**Probe**	**DOS**	**R2L**	**U2R**	**TOTAL**
Domain expert's Suggestion	**Normal**	4	0	0	0	3	7
	Probe	0	4	0	0	0	4
	DOS	0	1	3	0	0	4
	R2L	0	0	0	4	0	4
	U2R	0	0	0	0	1	1
	TOTAL	4	5	3	4	4	20

Table 6 Confusion matrix Domain expert vs system comparaison

	TP	**Precision**	**Recall**	**f-measure**	**Class**
	57	100	57	72.6	Normal
	100	80	100	89	Probe
	75	100	75	86	DOS
	100	100	100	100	R2L
	100	25	100	40	U2R
Weighted Average	86.4	81	86.4	77.5	

Table 7 performance evaluation (in %) based on precision, recall and F-measure

VIII. CONCLUSION

In the study, we integrated induced rules generated by employing JRIP with the knowledge based system resulting the rule based intrusion detection and advising KBS. Performance testing disclosed that the system has 80% of accuracy

with very good Precision and Recall. The performance analysis showed that RIDA-KBS registered acceptable result. In addition, the study has proven that possibility of updating both rule base and fact base of the knowledge base system whenever the data size changes. And then the knowledge base system also provides advice and information, based on the new changes yielding up-to-date knowledge base system. However, For the future, further exploration and study has to be done to refine and yield a better knowledge based system by optimizing rules which can be deployed in real network and provide advice to network administrators so that they can take timely and appropriate actions for a certain network incident.

REFERENCES

Abraham, A. (2005). Rule-Based expert systems. *Handbook of measuring system design*.

Barbara, D., Couto, J., Jajodia, S., Popyack, L., & Wu, N. (2001). *ADAM: Detecting intrusions by data mining.* Paper presented at the In Proceedings of the IEEE Workshop on Information Assurance and Security.

Bramer, M. (2007). *Principles of data mining* (Vol. 180): Springer.

Chauhan, H., Kumar, V., Pundir, S., & Pilli, E. S. (2013). *A comparative study of classification techniques for intrusion detection.* Paper presented at the Computational and Business Intelligence (ISCBI), 2013 International Symposium on.

Datta, R., & Saha, S. (2011). *An Empirical comparison of rule based classification techniques in medical databases*. Retrieved from

Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., & Tan, P.-N. (2002). *Data mining for network intrusion detection.* Paper presented at the Proc. NSF Workshop on Next Generation Data Mining.

Domingos, P. (2007). Toward knowledge-rich data mining. *Data Mining and Knowledge Discovery, 15*(1), 21-28.

Ektefa, M., Memar, S., Sidi, F., & Affendey, L. S. (2010). *Intrusion detection using data mining techniques.* Paper presented at the Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Ghorbani, A. A., Lu, Wei, Tavallaee, Mahbod. (2010). *Network Intrusion Detection and Prevention Concepts and Techniques*: Springer US.

Ilgun, K., Kemmerer, R. A., & Porras, P. A. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE transactions on software engineering, 21*(3), 181-199.

Jackson, J. (2002). Data Mining; A Conceptual Overview. *Communications of the Association for Information Systems, 8*(1), 19.

Kamber, J. H. a. M. (2006). *Jiawei Han and Micheline Kamber* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Krishnamoorthy, C., & Rajeev, S. (1996). *Artificial intelligence and expert systems for engineers* (Vol. 11): CRC press.

Lee, W., Stolfo, S. J., & Mok, K. W. (1999). *A data mining framework for building intrusion detection models.* Paper presented at the Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on.

Liao, H.-J., Lin, C.-H. R., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications, 36*(1), 16-24.

Mitchell, R., & Chen, R. (2013). Behavior-rule based intrusion detection systems for safety critical smart grid applications. *IEEE Transactions on Smart Grid, 4*(3), 1254-1263.

Modi, C., Patel, D., Borisaniya, B., Patel, H., Patel, A., & Rajarajan, M. (2013). A survey of intrusion detection techniques in cloud. *Journal of Network and Computer Applications, 36*(1), 42-57.

Oprea, M. (2006). On the Use of Data-Mining Techniques in Knowledge-Based Systems. *Economy Informatics, 6*, 21-24.

Shapiro, E. (1986). Concurrent Prolog: A progress report *Fundamentals of Artificial Intelligence* (pp. 277-313): Springer.

Srinivasulu, P., Nagaraju, D., Kumar, P. R., & Rao, K. N. (2009). Classifying the network intrusion attacks using data mining classification methods and their performance comparison. *International Journal of Computer Science and Network Security, 9*(6), 11-18.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A.-A. (2009). *A detailed analysis of the KDD CUP 99 data set.* Paper presented at the Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.

Tiwari, K. K., Tiwari, S., & Yadav, S. (2013). Intrusion detection using data mining techniques. *International Journal of Advanced Computer Technology, 2*(4), 21-25.

Turban, E., Aronson, J., & Liang, T.-P. (2005). *Decision Support Systems and Intelligent Systems 7"'Edition*: Pearson Prentice Hall.

Development of Knowledge Based System for Wheat Disease Diagnosis: A Rule Based Approach

Desalegn Aweke Wako[1], Million Meshesha (PhD) [2]
[1] Faculty of Computing, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia, desbdu@gmail.com
[2] School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia, meshe84@gmail.com

***Abstract* — Wheat production in Ethiopia is widely affected by diseases and attached by a number of insect pests. Wheat disease diagnosis needs sufficient and knowledgeable agricultural experts to identify the diseases and describe the methods of treatment and protection at early stage of infestation. But, agricultural specialist assistance may not always available and accessible to every farmer when the need arises for their help. Hence, this study presents an interactive rule based knowledge based system for wheat disease diagnosis in order to identify wheat diseases timely and apply the control measures effectively. The system aims to provide a guide for research centers and development agents to facilitate the diagnostic process. To develop the system, data and knowledge is acquired from documented and non- documented sources. The acquired knowledge is modeled using decision tree structure that represents concepts and procedures involved in the diagnosis of wheat disease. The system is developed using SWI Prolog programming language. The system has been tested and evaluated to ensure whether the performance of the system is accurate and the system is usable by research centers and development agents. The system has been registered the overall performance of 82%. So, the developed system has potential to use as a decision tool for diagnosing and treating wheat disease.**

Key words- **Agriculture, KBS, Rule based representation, Wheat, Disease**

I. INTRODUCTION

Agriculture is the mainstay of the Ethiopian economy and the livelihoods of more than 80% of the citizens (Diao et al., 2007). Ethiopia is the largest wheat producer in sub-Saharan Africa, next to South Africa. Wheat is the second most important in total production next to maize and the third in area after maize and sorghum that plays a significant role in assuring food sufficiency (Ethiopia Commodity Exchange Authority, 2008). Despite the expansion of wheat in most parts of Ethiopia, the country is not self-sufficient in production and consequently a large quantity of wheat is imported every year to fill the gap. The national average of wheat in the country, which is 14 qt/ha, is 24%, is still below the average of South Africa yield and 48% below that of the world's (Tesfaye et al., 2014). This low production & productivity is mainly due to diseases, frost drought, poor soil fertility, soil erosion, pests and problematic weeds (Hailu, 1991).

According to Hailu(1991), wheat production in Ethiopia is widely affected by diseases and pests. Such diseases and pests reduce yield, quality and marketability of the wheat crop. This problem needs sufficient and knowledgeable agricultural experts to identify the diseases and describe the methods of treatment and protection at early stage of infestation. Unluckily, agricultural specialist assistance may not always available and accessible to every farmer when the need arises for their help. Moreover, there is a shortage of extension agents to provide quick and timely decision making as each agent has to serve on average 1090 farmers (Belay & Abebaw, 2004). To address such problems the application of information technology is important to deploy in agricultural area. Knowledge based system is one aspect of information technology which provides the appropriate management for diagnosing diseases attacking crop (Ahmed, 2014) and (Prasad & Babu, 2006). Knowledge Based System (KBS) is computer program that replicate the reasoning processes of a human expert in order to deliver a solution concerning a problem (Sajja & Akerkar, 2010).

Hence, this study presents the development of rule based knowledge based system for wheat disease diagnosis that can provide advice for research centers and development agents to facilitate the diagnostic process. KBS can act as an expert on demand without wasting time, anytime and anywhere. With the proper utilization of knowledge, the knowledge based systems make decisions, recommendations and perform tasks based on user input, increase productivity and document rare knowledge by capturing scarce expertise and enhances problem solving capabilities in most flexible way (Sajja & Akerkar, 2010).

II. RELATED WORKS

There are many works in the literature that explains about knowledge based systems in the agriculture domain. Local research work shows that Berhanu (2012) developed a knowledge based system for coffee disease diagnosis and treatment. A knowledge based system for cereal crop diagnosis and treatment is explored by Ejigu (2012). The

focus of the study was to address problems of common diseases occurring in cereal crop.

Moreover, internationally Hogeveen et al. (1991) developed an integrated knowledge based system for dairy farms that serves diagnostic, problem solving, and advising role in controlling the health and production of a herd, including the financial consequences. King et al. (1991) developed a knowledge based system for malting barley management. The system gives advice on fertilizer and water applications to maximize crop yield under strict quality constraints. Yelapure et al. (2011) developed knowledge based system for tomato crop with special reference to pesticides. This system helps to identify the pests and to suggest pesticide treatment to control it.

To conclude, several studies have been developed in AI using knowledge based systems to reason out the solution of a particular problem. But, according to the researcher's knowledge, there are no research conducted to design a knowledge based system for wheat disease diagnosis and treatment. Thus, in this study an attempt is made to develop a rule based knowledge based system for wheat disease diagnosis and to test and evaluate its performance with the help of professional experts in the field.

Therefore, the proposed knowledge based system can assist domain experts during wheat disease diagnosis and treatments by providing advisory services. This work will also be used as an input for further disease diagnosing and management in agricultural industry.

III. METHODOLOGY OF THE STUDY

In this study, different procedures are followed in developing the proposed knowledge based system. These are: knowledge acquisition, knowledge modeling, knowledge Representation, Knowledge based system development for wheat disease diagnosis and Evaluation of the system.

The primary knowledge needed for this study is acquired from Debre zeit Agricultural research center. Four domain experts are selected using purposive sampling techniques and interviewed to extract the tacit knowledge. Similarly, documented sources of knowledge are consulted on the area of crop protection and treatment from different sources such as agricultural books, journals, publications, internet sources, plant disease protection guidelines and training manuals are analyzed.

The acquired knowledge is modeled using decision tree. Decision tree shows the relationships of the problem graphically and can handle complex situations in a compact form. Knowledge diagramming is often more natural to experts than formal representation methods and decision trees can easily be converted to rules. Decision tree is drawn using flow chart symbols as it is easier for many to read and understand. It helps to identify a strategy most likely to reach a goal and allow the addition of new scenarios

After modeling the acquired knowledge using decision tree, it is represented in a format that is both understandable by humans and executable on computers. Production rules are the most popular form of knowledge representation which is easy to understand and reasonably efficient in diagnosing problems. Knowledge is represented in the form of **condition-action** pairs: **IF** this condition (or premise or antecedent) occurs, **THEN** some action (or result or conclusion or consequence) will (or should) occur.

Prolog programming language is used to develop a rule based knowledge based system for wheat disease diagnosis. The reasons for selecting Prolog are the features and abilities of the language that incorporate it. Prolog is a declarative language and has the capacity to describe the real world. Because of its declarative semantics, built-in search, and pattern matching, Prolog provides an important tool for programs that process natural language.

To achieve the established objective of the study, the prototype system is extensively tested and evaluated to ensure that whether the performance of the system is accurate and the system is usable by research centers and development agents.

IV. MODEL DEVELOPMENT

Wheat disease diagnosis KBS is designed based on rule based reasoning techniques. As shown in Fig.1 below, the model shows that, the knowledge is acquired from experts and documented knowledge sources. Potential sources of knowledge include domain experts, books, journal articles, proceedings, electronic sources and information available on the web. Then the acquired knowledge is effectively coded in the knowledge base by knowledge engineer. Knowledge base contains rule base from which the system draws conclusion through inference engine. The inference engine accepts query from the user via user interface and prompt the action in user understandable form if the goal is satisfied. A backward chaining technique is used as inference mechanism to search and extract the rules for specific type of wheat disease.

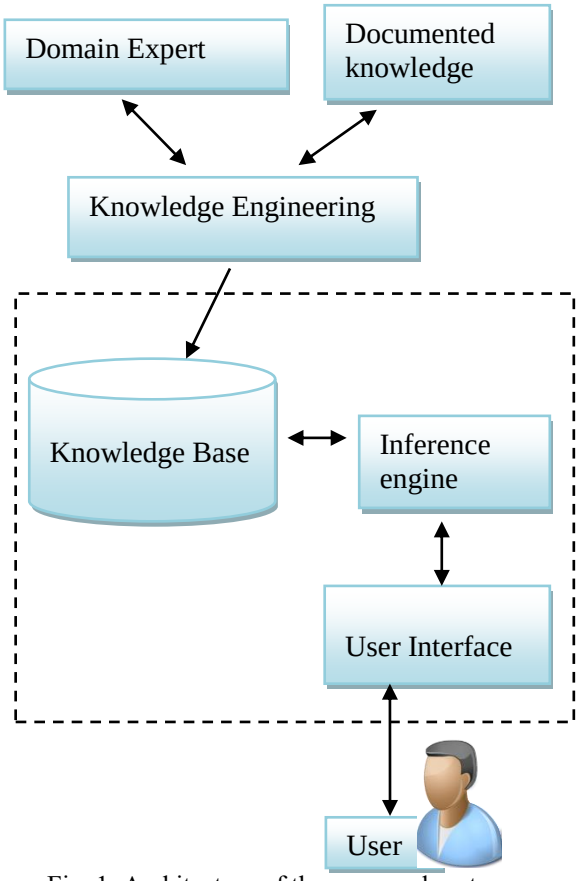


Fig. 1. Architecture of the proposed system

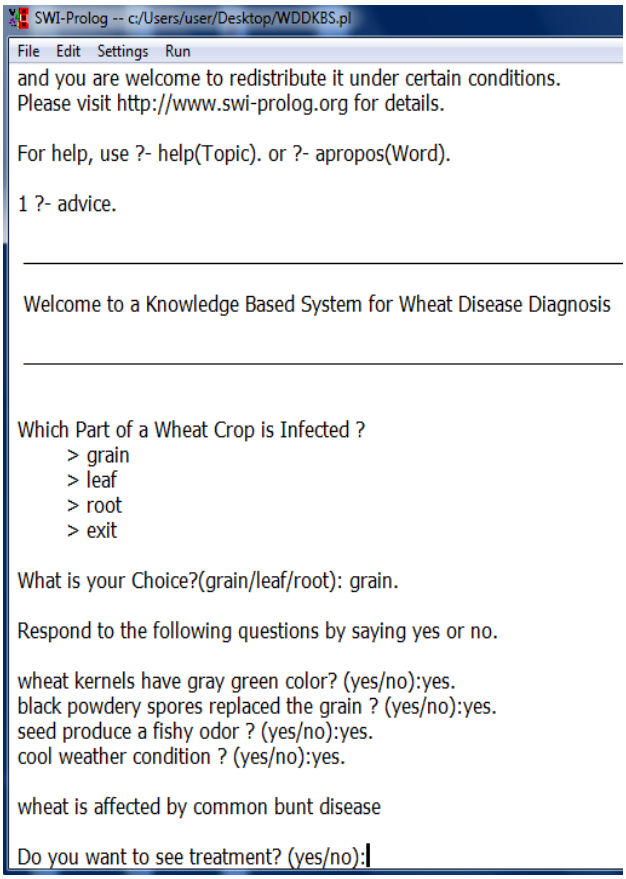


Fig. 2. Sample dialogue of the proposed system during diagnosis

V. PROPOSED SYSTEM USER INTERFACE

To interact with the system the user interface is needed. User interface is a bidirectional communication between the system and the user. It is the window through which the system is able to return information to the user

The end-users can start the system by opening the file 'WDDKBS.pl' and consult the system for diagnosis by typing the word "advice" followed by full stop(.) . Then the welcome window is displayed automatically along with the menu containing choices as shown in Fig.2. To enable the system to diagnose the wheat disease, the system requests the end-users to choose the infected part of the wheat. If a choice match, the system can diagnose the disease by asking the most common signs and symptoms.

VI. SYSTEM TESTING AND EVALUATION

System testing and evaluation of the prototype system is the final step that assists the knowledge engineer to measure whether the prototype system is met the proposed objective or not.

System testing involves testing the rule based reasoning modules in order to determine the performance of the proposed system. More importantly, evaluation is carried out to determine users' acceptance and applicability of the prototype knowledge based system in the domain area. The purpose of the evaluation is to get the end user's views on the significance or usefulness of the system. The evaluation and testing issue of the system answers the question "To what extent the proposed system gives acceptable and accurate diagnosis and treatment to wheat disease?" To answer this question, system performance testing and user acceptance testing methods are used.

A. User Acceptance Testing

User acceptance testing is the process of ensuring whether the system satisfies the requirements of its end-users. A questionnaire is prepared to evaluate the user acceptance of a system, and the evaluators fill the questionnaire after they have used the system. The researcher adopted the questions from

Mbogho (2012) and Solomon (2013) represents the questions that users used to evaluate the system. Table II shows the evaluation results. The evaluators fill in the questionnaire as excellent, very good, good, fair and poor for each of the questions. The author assigned values for each word as excellent = 5, very good= 4, good= 3, fair=2 and poor = 1. A total of four domain experts are participated in the system evaluation.

Query number	Questions
1	Simplicity to use and interact with the system
2	The system operates in a very good speed and efficiency
3	The accuracy of the system in reaching a decision to identify wheat disease
4	Coverage of domain knowledge is sufficient
5	The ability of the prototype system in making right conclusions and recommendations
6	Contribution of the prototype system in the domain area

Table I. Evaluation Questions for user acceptance testing

The user acceptance of the system is computed manually using Equation (1).

$$AveS = SV1*\frac{nr1}{tnr} + SV2*\frac{nr2}{tnr} + ...\sum_{i=1}^{n} SVi*nri/tnr \quad (1)$$

Where, AveS average score, SV scale value, TNR total number of respondent and NR is number of respondent. To get the result of user acceptance average performance is calculated out 100%.

$$Avp == AveS/NS *100 \quad (2)$$

Where, NS is number of scale and Avp is average performance.

Query	Excellent	V. good	Good	Fair	Poor	Average Score	Average Performance (%)
1	1	1	1	1	0	3.50	70
2	2	1	1	0	0	4.25	85
3	1	2	1	0	0	4.00	80
4	2	2	0	0	0	4.50	90
5	2	1	1	0	0	4.25	85
6	3	1	0	0	0	4.75	95
	Total Average					4.21	84.2

Table II: Results of Evaluation based on the evaluation questions

The average score of each questionnaire is calculated using the sum of values of excellent, very good, good, fair and poor and divided the sum by four as illustrated in equation (1). Table II shows that 25% of the evaluators scored the simplicity to use and interact with the system criteria of evaluation as excellent, 25% as very good, 25% as good, and 25% as fair. One evaluator rate the simplicity to use and interact with the system as fair. The reason behind is that the evaluator wants to retrieve queries via user interface in his local language and wants the system to provide decisions in his local language, so as to understand the decisions made by the system. The second evaluation criteria efficiency in time showed a greater rate of efficiency by the evaluators, in which 50% rated as excellent, 25% as very good, and the rest 25% as good.

50% of the evaluators gave the prototype system a very good score with regard to the accuracy of the prototype system in reaching a decision to identify the wheat disease, 25% as excellent, and 25% as good. And when asked if the prototype system included adequate knowledge to diagnose and treat wheat disease, 50% of the evaluators rated the prototype system as very good and 50% rated as excellent. The ability of the prototype system in making right conclusions and recommendations criteria was scored by the evaluators 50% as excellent while 25% as very good, and 25% of the evaluators scored it with good.

Similarly, the final evaluation criteria concerning the system contribution in the domain area, 25% of the evaluators gave the prototype system very good while 75% rated the prototype system as excellent. As a result, based on the responses of four system evaluators, the average performance obtained is 4.21 on scale of 5. This value is the result obtained from the values assigned for each close ended question. The result indicates that about 84.2% of evaluators are satisfied by the performance of the knowledge based system. This implies that the developed prototype system performs well in making right advice on diagnosing and treating wheat disease.

B. Performance Testing

The performance of the system is tested and validated using test cases. The test cases are used to measure the accuracy of the prototype system. For the test, a total of fifteen test cases are selected. Test cases that have similar parameters with the prototype system are selected purposively. These test cases are categorized into three based on their resemblances and in their characteristics. These are generalized as root disease cases, leaf disease cases and grain disease cases.

Additionally, four domain experts are selected to evaluate the system by interacting with the developed system. These evaluators have participated in the user acceptance evaluation. The reason why domain experts are selected for the second time was that they are familiar with the

system during visual interaction evaluation. So they can easily understand about the system. The testing procedure is carried out by system evaluators to classify the test cases into correctly or incorrectly diagnosed cases. System evaluators compare the decisions made by the system with that of the experts' decision on those cases. Then system evaluators validate the numbers of correct decisions made by the system. The result of the comparison shows that the rule based system has made close decision in the process of diagnosing wheat disease as human expert do. As indicated in table III below, the test case result provided by system evaluators showed that the proposed knowledge based system is about 80% accurate in diagnosing wheat disease.

Selected Cases	# of cases selected for testing	CDC	IC.DC	Accuracy
root disease cases	5	4	1	80%
leaf disease cases	5	3	2	60%
grain disease cases	5	5	0	100%
Total	15	12	3	80%

Table III. Performance testing

In the above table III, few abbreviations are used. CDC stands for correctly diagnosed cases, IC.DC stands for incorrectly diagnosed cases. Fifteen test cases are selected purposively to validate the accuracy of the system and five cases are assigned for each selected disease cases. As a result, for root disease cases four of them are correctly diagnosed from the given five cases. Similarly, from the given five cases again three of them are diagnosed correctly for leaf disease cases. Finally, the system classified all of the given cases for grain disease cases and it achieves the maximum performance. The result indicated that all the cases are directly similar with knowledge incorporated in the knowledge base.

As discussed in section VI under subsections *A* and *B*, the average evaluation result filled by the domain experts in the domain area is 84.2% and the accuracy of the prototype system is calculated as 80% respectively. The overall performance of the prototype system is 82%.

VII. RESULTS AND DISCUSSION

In this study both system performance testing and user acceptance testing have been done for the prototype knowledge based system. In measuring the performance (accuracy) of the system, rule based knowledge based is validated against the specified cases and the accuracy of the prototype system is calculated as 80%. In addition to accuracy, user acceptance evaluation of the prototype system has been calculated as 84.2%. The overall performance of the prototype knowledge based system is 82%.

Depending on the results found the main strength of the prototype system in the domain areas are:

- The system is promising and helps to solve problems in the areas where experienced and skilled agriculture experts are unavailable.
- The system is very helpful to solve problems timely with accumulated knowledge.
- The system can reduce the existed knowledge gap observed in remote areas where skilled agricultural professionals are not available.
- The system helps as a knowledge sharing and training tool for DA (development agent) workers, it can improve the skill of DA workers in wheat disease identification and decision making.

Regardless of the strength of the system, the researchers have faced some challenges during the study which limits the system to register a better performance for diagnoses of wheat disease. These are discussed as follows:

- Evaluators want to see the severity of the infestation levels of the affected wheat crop
- Users who lack computer skills and access might not implement it.
- The system needs to incorporate multiple languages to respond in their local language.
- The performance of the prototype system depends directly on the quality of the knowledge acquired from domain experts. However, knowledge elicitation from domain experts are the most difficult task due to the tacit nature of persons, knowledge is difficult to transfer to another person by means of writing it down.
- Language barriers and lack of adopted KBS technology in agricultural industry in Ethiopia hinders the system not to register high performance.

VIII. CONCLUSION AND FUTURE WORK

In this paper, an interactive rule based knowledge based system was developed to diagnose wheat diseases in order to solve problems in the areas where experienced and skilled agricultural experts are unavailable. The system provides advisory service for research centers and development agents to facilitate the diagnostic process. The system was

evaluated using different evaluation methods and achieved 82% of the overall performance. The prototype system achieves a good performance and meets the objectives of the study. However, in order to make the system applicable in the domain area for diagnosis of wheat disease, the user interface should support local languages to meet the needs of local users and more research work must be done to incorporate high quality pictures which depict the symptoms in order to identify the damage level of the diseased part of wheat crop.

The proposed tool has potential to use as a decision tool for diagnosing and treating wheat disease. However, it still needs a future validation with more cases. In addition, expert decision is regularly hard to measure with precise numerical data. So in the future, fuzzy set theory will need to be integrated in to a proposed knowledge based system.

REFERENCES

Ahmed, R. (2014). *Expert System Applications: Agriculture.* Retrieved January 05, 2017, from www.arc.sci.eg/narims_upload/claesfiles/3759.pdf

Belay, K., & Abebaw, D. (2004). Challenges Facing Agricultural Extension Agents: Case Study from South-western Ethiopia. *African Development Bank 2004* , 139-168.

Berhanu, A. (2012). Developing a Knowledge based system for coffee disease diagnosis and treatment. Addis Abeba,Ethiopia: MSc Thesis,Addis Ababa University, School of Information Science.

Diao, X., Belay, F., Haggblade, S., Alemayehu, S., & Kassu, W. (2007). Agricultural growth linkages in Ethiopia: Estimates using Fixed and Flexible Price Models. 1-41.

Ejigu, T. (2012). Developing knowledge based system for cereal crop diagnosis and treatment. MSc Thesis, Addis Ababa University, School of Information Science.

Ethiopia Commodity Exchange Authority. (2008). *Understanding Wheat: A Review of Supply and Marketing issues.* Addis Abeba, Ethiopia.

Hailu, G. (1991). Use of germplasm resources in breeding wheat for disease resistance. 298-302. InEagles JMM, Hawkes JG, Worede M (eds): Cambridge University Press.

Hogeveen, H., Noordhuizekstassen, E. N., Schreinemakers, J. F., & Brand, A. (1991). Our Industry Today Development of an Integrated Knowledge-Based System for Management Support on Dairy Farms. *Journal of Dairy Science , 74* (12).

King, J., Broner, I., L, C. R., & Basham, C. (1991). Malting Barley Water and nutrient Management Knowledge-Based System. Emerging Technologies, Div.of ASAE .

Mbogho, A. (2012). Knowledge Based System for Medical Advice provision. Cape Town,South Africa: MSc thesis, department of Computer Science, University of Cape Town.

Prasad, G., & Babu, A. V. (2006). A Study on Various Expert Systems in Agriculture. Georgian Electronic Scientific Journal:Computer Science and Telecommunications (4(11)), 81-86.

Sajja, P. S., & Akerkar, R. (2010). Knowledge-Based Systems for Development. Advanced Knowledge Based Systems: Model, Applications & Research , 1, 1-11.

Solomon, G. (2013). A Self Learning Knowledge Based System For Diagnosis and Treatment of Diabetes. Addis Abeba,Ethiopia: MSc Thesis,Addis Abeba University,School of Information Science.

Tesfaye, T., Genet, T., & Desalegn, T. (2014). Genetic variability, heritability and genetic diversity of bread wheat (Triticum aestivum L.) genotype in western Amhara region. Wudpecker Journal of Agricultural Research , 3(1), 026-034.

Yelapure, S. J., Jadhav, S. K., & Babar, V. M. (2011). Knowledge based System for Tomato Crop with Special Reference to Pesticides. International Journal of Computer Applications (IJCA) 25 .

Ontology Development for Anemic Pregnant Women

Mamo Abebe [1], Esubalew Alemneh [2]

[1] Faculty of Computing, Bahir Dar Institute of Technology, Bahiar Dar University, Bahiar Dar, Ethiopia, mamoa2009@gmail.com

[2] Faculty of Computing, Bahir Dar Institute of Technology, Bahiar Dar University, Bahiar Dar, Ethiopia, esubalew@gmail.com

***Abstract*- Health care models enable patients with anemic diseases to acquire continuous and long-term medical services at home or at the back of the doctors. This improves health care and delivery of medical services can be accessed anywhere. Today high prevalence of anemia diseases will pose technological challenges in pregnant women's health care. This research investigates on how ontology in semantic technologies could address the above challenges. Ontology enables to transfer implicitly expert knowledge to explicit ontology engineer during model development. The goal of this paper is to use semantic technology for building ontology-based knowledge repository that provides a model which helps pregnant women to diagnose themselves. By inputting signs, symptoms and feeding habits of pregnant woman to the model a pregnant woman can check her about anemia and get recommendations for anemia diagnosis. The ontology depicts the conceptual representation of the domain of anemia disease diagnosis, along with data sources and corresponding mappings allows us to translate the SPARQL queries into the data level queries and executed by the underlying database management model. Based on ontology's structure, the model can collect, store and share information from heterogeneous sources and reason over the knowledge.**

Ontology-based health care model is a novel decision support model for healthcare services that supports clinical decisions for anemic disease. The model contains knowledge about symptoms, sign and medication of diseases in pregnant women appearing during their lifespan of pregnancies. Medical officers who involve directly with pregnant women may use this model to assist them in managing the anemia disease control. For development purposes, knowledge engineering methodology is selected as a guide. Perhaps, this model will become the most popular alternative for its performance and guide healthy pregnant women to be free from anemia diseases. The model is evaluated using visual, interactive methods; it is shown that the model agrees with human expert. The average evaluation result filled with the domain experts in the domain area is 90.3%

Key words- **anemia ontology, ontology, ontology engineering, knowledge acquisition.**

I. INTRODUCTION

Anemia is a global public health problem both in developing and developed countries affecting people of different age groups. However, it is more prominent in pregnant women, young children and other reproductive age. According to the 2008 WHO report, anemia affected 1.62 billion (24.8%) people globally. It had an estimated global prevalence of 42% in pregnant women and was a major cause of maternal mortality. Sub-Saharan Africa is the most affected region, with anemia prevalence estimated to be 17.2 million pregnant women, which corresponds to approximately 30% of total global cases. According to (WHO,2015) "In developing countries, every second pregnant woman and about 40% of preschool children are estimated to be anaemic" (Ejeta et al., 2014). Ethiopia is among countries where there is a high level of anemia among women of reproductive age (15-49 years) and pregnant women. A higher proportion of pregnant women are anemic (22%) than breastfeeding women (19 %) and women who are neither pregnant nor breastfeeding (15 %). It is estimated that more than 2 billion people are iron deficient globally. Among these people 1.2 billion become severely anemic. About 90% of all anemias have an iron deficiency component. In the developing world, nearly half of the population has iron deficiency. However, the industrial world is not free from it; 11% of its population have iron deficiency (TSEHAYU, 2009).Iron deficiency anemia was ranked as one of the significant micronutrient deficiency problems in Ethiopia. As study conducted in Ethiopia prevalence of anemia in pregnant women: In jiniga at Karamar hospital 52.9% are anemic (Alemnesh ,2013), In Somali region 46.8% are anemic (EDHS ,2011), Southern Ethiopia at Boditii Health Center 61.6% are anemia, In the Gilgal Gibe dam area 53.9% are anemic, and in Sidama 31.6%, in Wester Arsi 36.6%, in wester Tigray 36.1% (Gebre,2015), in Asendabo (62.7%) anemic (Lelissa, 2015), in Jimma 57%, in Harar 27.9%, anemic, In Gondar (23.2%) and Bahiar Dar 13% (matekaki,2016). All of them are conducted study on prevalence of anemia. The prevalence of anemia is currently increased and the communities affected by nutrition, hookworm, low income status, malaria, parasitic infection. Blood loss also contributed to anemia disease. Therefore, we need additional ontology-based semantic web applications that can recommend and diagnosis

anemia. There are many challenges that make the diagnosis of anemia in pregnant women difficult and sometimes impossible to progress further in the direction. The problem that the knowledge acquired through experience doesn't get reused because it isn't shared in a formal way. The challenge of non-availability of a specialist for diagnosis of anemia. The challenge of non-availability of health centers in proximities and the necessity of regulating and preventing anemia among the pregnant women. Therefore, to improve the aforementioned problems we have developed an ontology-based decision support model to diagnose anemia in pregnant women

1.1. Ontology

The concept of the ontology comes from the realm of philosophy. In the 1960 s, the field of computer began to use ontology, but until 1993 Gruber gave the ontology standard definition: ontology is the clear specification of the conceptual model (wright, 1992). At present, the definition of ontology is the clear formal specification of the sharing the conceptual model. It can be taking as four levels, such as the conceptual model, clarity, formalization and share. Specific description as follows: The concept model, obtained by abstracting out some of the objective world phenomenon related concepts. The meaning exhibited by concept model is independent of specific conservational state; Clarity, means that the concept used and constraints used by concepts are clearly defined; Formalization, with the body language code, ontology can make the computer be read and processed. Share, ontology embodies knowledge recognized commonly and reflects the concept set recognized publicly in the relevant fields, that is, ontology aimed at groups rather than individual's consensus.

In general, the goal of onto1ogy is to obtain, describe and express knowledge in related fields, and provide the common understanding of the domain knowledge and make the vocabulary knowledge clearly recognized commonly in the field, and give the definition between these words and relationship between these concepts/objects clearly. General speaking, the onto1ogy is a file which officially define a group of words and the relationship between them. It is the integration of the knowledge representation and the relationship rules for intelligent model. It defines some relevant classes and objects to express the knowledge. The relationship rules express the relationship between knowledge and provide a kind of intelligent reasoning mechanism

1.2. Significance of the Study

The substantial significance of the study is to better inform pregnant women, directly or indirectly, about anemia diseases that threaten pregnant women in Ethiopia, around whom there are less number of experts to help them. This will better empower them to take action as required. If a significant association is present, the study will recommend of developing a protective nutritional and treatment plan during pregnancy to reduce the risk development of anemia in pregnant women. To overcome the problem aforementioned, diagnosis of anemia disease is important for pregnant women, and help non-experts in identification of sign and symptoms of anemia diagnosis. Non-expert can improve their knowledge and be effective in anemia disease identification and diagnostic work as expert knowledge is readily available. Expert knowledge is not long lived, but when expert knowledge is coded or captured by help of ontology, it is long live in community. Till now there was no ontology base model for pregnant women diagnosis, so our model provides important role in community to diagnose anemia remotely or at home by themselves. The model provide service for dietary, socio-demographic, health status, health survives and mother and child history-diagnosis and recommendation of dietary taken during each trimester (1-9 month). The model convinced whether pregnant women anemic or non-anemic. Made simple in accessing best physician knowledge because the model is modeled from the best expert in the area, saves time and cost in exploration of physician for treatment

II. RELATED WORKS

Cancer is a class of diseases characterized by out-of-control cell growth. There are over 200 different types of cancer, and each is classified by the type of cell that is initially affected. They proposed medical systems for cancer diseases. It also proposes an ontology-based diagnostic methodology for cancer diseases. This methodology can be applied to help patients, students and physicians decide what cancer type, the patient has, what is the stage of the cancer and how it can be treated (M. ALFONSE, M., AREF, M., & SALEM).The Breast Cancer ontology models the knowledge encapsulated within the breast cancer follow-up clinical practice guideline. They used the Protégé ontology editing environment to build their Breast Cancer ontology in OWL. Given a series of conditional recommendations, the first challenge in the ontology engineering process was to identify the concept classes, their properties, the decision variables dictating the choice of the

recommendations logical operations in the recommendations (M . Alfonse, Mostafa , Aref, Badeeh, & Salem, 2001).

Diabetic1 is the leading cause of death in many countries. Surgery in diabetic patients is more complicated than nondiabetic patients. Severe hypoglycemia can increase the mortality rate and cardiovascular deaths. In order to address the issues, they adopted structured approaches such as clinical guidelines which are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances. Clinicians can use the guidelines to standardize clinical practice and ultimately to improve the outcomes of patients through an Ontological Approach for Guideline-based Decision Support System. However, reference indicated that medical guidelines are underutilized at the point of care. The issues can be tackled by providing decision making to assist clinicians with an ontology-driven clinical decision support system (CDSS). The development tool they used was OWL, RDF, SPAQL, JENA, Fuzzy Logic, protégé for the addition of ontology and class they use disease, anesthesia, medication. The proposed ontological-approach took account of one guideline and its corresponding domain ontology to establish the guideline execution system (Chen & Bau, 2013).

Diabetic2: a diagnosis and treatment recommendation system for diabetes. The system considers patient information, symptoms and signs, risk factors and lab tests and suggests a treatment plan according to the diabetes type as recommended by the CPG. The development procedure consisted in the acquisition, modeling and implementation of diabetes domain expertise from experts, the CPG and other sources to develop domain ontology and a decision support system to handle the diabetes in an early stage. The proposed system uses ontology to allow a standard representation of domain concepts and relationships and enable clinical knowledge sharing, update and reuse. The proposed ontology is designed and developed by OWL-DL, the rules are constructed by the SWRL and executed by JESS inference engine (Alharbi, Berri, & Masri, 2015) We are developed anemia ontology using the tools and methodology what other used. Comparing of our model with related works following table

2. Methodology

The Methodology provides information that is involved in cooperation in understanding concepts, theories and principles for conducting and developing a research design. To undertake this research work, the following methods have been used. The methodology is used to build ontology base decision support model to diagnose anemia in pregnant women. At the outset, intensive literature review is made. Then domain and scope of ontology development are defined and domain specific interest needs are collected from relevant literatures. After a series of sorting and refining, these concepts are used as bricks to build the ontology. Finally, a hierarchical structure, together with the properties of its multi-level subclasses is generated on the basis of terms of concepts. Relevant case studies, article and thesis paper are reviewed to check the comprehensiveness as well as representations of the ontology.

The study is based on anemia diseases encountered on pregnant women in Ethiopia. Expert knowledge is not enough to diagnose anemia manually because of a limited number of experts or clinicians to treat pregnant women during pregnancy, but now the model we built can overcome this problem of dearth of experts and asylums expert knowledge for all in the relevant field. The knowledge behind creating a knowledgeable model can enable many people to be benefited from the knowledge of a domain specific expert. Expert model simulates the judgment and behavior of a human that has expert knowledge and experience in a particular field. In an expert model development, ontology base development is the most important part. The quality of an expert model depends on its ontology base. Ontology-based development with the help of domain specific expert in the model is developed through an ontology editing tool called protégé.

The process of developing ODAPW has a multi-step process of developing a domain-specific ontology base. The ladders for developing our ontology-based decision support model passes through different ontology modeling design stages: First, we identify the input problem in our area of the domain. The second stage comprises of ontology (knowledge) acquisition. Third phase involves model design methodology and developing the process. Fourth stage encompasses the knowledge modeling and representation of knowledge into the knowledge base. Next, we map the database to ontology in the fifth stage. At the end, we establish various production rules. A methodology of ontology design describes all activities necessary for the construction of ODAPW.

2.1. Knowledge Acquisition

Knowledge acquisition is the process used to define the rules and ontologies required for a knowledge-based system. The phrase was first used in combination with expert systems to describe the initial tasks associated with developing an expert system, namely finding and

interviewing domain experts and capturing their knowledge via rules, objects, and frame-based ontologies. Expert systems were one of the first successful applications of artificial intelligence technology to real world health problems. Knowledge acquisition is the process of transferring knowledge from the knowledge source to knowledge engineer then encoding it into the knowledge base(Potter; Roussey, 2005). The sources of knowledge in our model are primarily experts. Secondly, literatures are consulted to acquire the knowledge. Domain ontology is obtained through individual interviews with experts in nutrition and symptom pathology. Moreover, we review published research articles and anemia disease management guides for the purpose. Basically, structured interviews are used to acquire knowledge of experts in the domain of the study; this is because of a structured interview is a systematic and goal-oriented process. It forces organized communication between the ontology engineer and the expert. The structure reduces the interpretation problems inherent in unstructured interviews and allows the ontology engineer to prevent the distortion caused by the subjectivity of the domain expert (Boose & Gaines, 1989). Therefore, interpersonal communication and analytic skills are important to come up with the quality knowledge base. We gain the knowledge of the various incursions from specific literature and symptom descriptions and the rules from domain experts. Two stages are iteratively done to implement this prototype. The first step is the domain acquaintance, during which, the ontology engineer targets to characterize the key problems and becomes familiar with the domain. This is accomplished by studying texts and articles relating to anemia and identifying appropriate domain experts to be consulted. The second step is the consulting with the experts, during which, we consult with the domain experts. The interviews with domain experts provide a lot of help in getting the idea of the extent of knowledge required to solve the problems. The experts are asked mainly the following questions: “what is an association factor of anemia?”, “What is causing anemia?”, “What has mostly been a symptom of anemia that affects pregnant women during pregnancy in the country?”, “Mainly what is the effect of anemia during pregnancy?”, and “What are the symptoms of malaria which cause anemia?” And “what are the basic treatments and preventions to be taken?” Initially, we make the effort to identify the disease, symptoms and treatments.

Data collection is conducted by organizing interview questionnaire. The questionnaire mainly focuses on socio-demographic, malaria, mother and child history, health status, dietary intake and health survives factors. Socio-demographic information, present and past history in pregnant women, environmental related factors and dietary habit, Plasmodium infection prevalence and concentration are assessed from the history of secondary data. We trained how to gather data in data collection procedure for this particular study to attain standardization and maximize interviewer reliability. The data collectors are regularly supervised by the principal investigator for proper data collection. Then, training the model is designed.

2.2. Ontology Engineering

Ontology engineering is a successor of knowledge engineering, which is considered as a key technology for building knowledge-intensive model. Although knowledge engineering contributes to elicit expertise, knowledge, organize it into a computational structure and then build knowledge bases, AI researchers have noticed the necessity of a more robust and theoretically sound engineering field which enables knowledge sharing/reuse and formulates the problem-solving process itself. To do this, it is fruitful for an ontology builder to answer several questions: What are the main components of ontology to be built? How does ontology look like and how does it work? Confirm if it is required to consider reusing the existing ontologies or not. What is the difficulty of the ontology to be developed? What are the principles of ontology design and development? How to evaluate ontology? And how is data collected? All these aforementioned ontology engineering acquires knowledge from domain experts through knowledge acquisition (Studer, Abecker, & Grimm, 2007).

2.3. Ontology Engineering as Modeling Process

Ontological engineering plays an important role in developing or building ontological model which diagnoses anemia in pregnant women, and ontology engineers are its practitioners. Ontology engineering is an applied part of the science of artificial intelligence which, in turn, is a part of computer science. Theoretically, an ontology engineer is a computer scientist who knows how to design and implement programs or model that incorporate artificial intelligence techniques.

2.4. Conceptual Knowledge Modeling

During the knowledge modeling stage, the expert’s knowledge (elicited by various techniques) is represented in a knowledge model. A knowledge

model is a systematized representation of knowledge using symbols to represent bits of knowledge and the relationships between them. Knowledge models include symbolic character-based languages such as propositional logic, first order logic and descriptive and tabular representation such as matrices and structured text like hypertext. The generation and modification cycle of a knowledge model is an essential part of the knowledge modeling phase. The model helps to ensure that all pregnant women understand the language and terminology being used and quickly conveys information for validation and modification where necessary. The knowledge models are also of great value during cross-validation with other specialists (Emberey & Vermeulen, 2007).

2.5. Model Design Methodology

To develop an ontology-based decision support model to diagnose anemia in pregnant women, first we need to identify the problem and understand the major characteristics of the problem under our domain that we have to solve in the ontology-based decision support model. The input to our model mainly is questions and answers as well as symptoms and signs to diagnose anemia in the pregnant women occurring during their pregnancy time. Figure3.1 illustrates our ontological model development process.

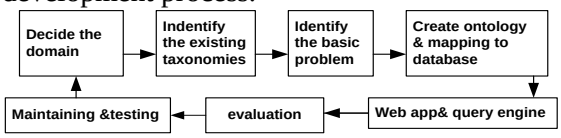
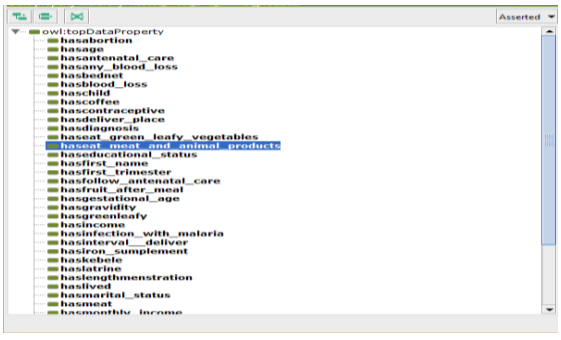


Figure3.1: methodology for developing anemia ontology

2.6. Knowledge Representation and Knowledge Model Based on ontology

Aiming is to identify anemia disease complex factors, such as the symptoms of anemia disease, pathogenesis regularity complex, vulnerable to environment condition, human factors and the difference among areas; it is difficult to use an only human knowledge representation to obtain accurate expression. Taking use of ontology, some bottleneck problems in the medical expert system, including the knowledge representation, knowledge organization and knowledge sharing, can be well solved. The goal of ontology is to obtain, describe and express the knowledge of the related field, provide the common understanding of the domain knowledge, determine the common admissive vocabularies in the field and give a clear definition of the relationship between these vocabulary words and terms from different levels of formal model. Based on above aims, this subject set up the database of anemia disease ontology. We adopt Protege5.0 to construct the anemia disease ontology database, use seven main concepts in the constructing process, take formalism description on target ontology and generated the owl file (S. Staab, studer,R., 2001). Specific steps as follows: List the important terms of ontology: There are many concerned terms in describing anemia disease. In order to extract the corresponding information, we choose some necessary terms, such as disease common symptom, disease sign, mother and child history, feeding habit, sociodemographic malaria, hookworm and prevention-control method ontology of the desired subject is defined as creating first and most striking concepts and making generalizations based on state privatizations. In this study, since classes of ontology to be created are in a systematic list from upside to down, top-down method was considered appropriate to proceed faster. If list of classes to be created is defined completely, using the general to the particular method provides convenience to the developer(Pinto & Martins, 2000). A datatype property is different from object property. Datatype property and object property, describe what kind of values a triple with the property should have. Datatype properties relate individuals to a literal data (e.g. Strings, numbers, date types, Boolean, etc.), whereas object properties relate individuals to other individuals. Something like hasAge would typically be a datatype property, for an age is a number, but HasDisease would be an object property, since a disease is another individual. Defining the properties of classes, slots. Classes and hierarchical structure created; on their own do not show clearly the information to be given to the desired audience. Semantic relations to specify the characteristics of properties and class definitions can



be used in the ontology.

FIGURE3. 2 OBJECT PROPERTY OF ANEMIA ONTOLOGY

There are two kinds of properties; Object properties and data properties. Object properties defined already created a relationship between two classes, or external parts and the characteristics of the class (Josep Blat, Ibáñez, & Navarrete).Within

Protégé atmosphere the whole picture of ontology domain is described into classes and class hierarchy. Classes provide an abstraction mechanism for, grouping resources with similar characteristics. It specifies concept of the domain as a collection of abstract objects defined by the same values of aspects. In ontology design, we use OWL language. Every OWL class is associated with a set of individual

Person class contains information about human which is categorized in this model development, such as pregnant women and healthcare providers. Each pregnant woman is also linked with the diseases class, symptom/sign classes by means of different relationship. This class will contain all types of individual's pregnant women who will interact with model; it also contains all necessary information used to describe each person such as pregnant woman's name, address, gender, age.

Symptom/sign class is class contains information about what type of symptom/sign we are describing. As our many focus for this model is anemia symptom, we describe anemia symptom, the individuals of this class have a direct relationship with pregnant women and the corresponding sign/symptom. Recommendation Class is a superclass of nutrition, medication classes, these classes provide information called context data. Those are the factors that can influence the variety of HBR of anemic pregnant women both negatively and positively including drugs, foods and drinks. The individuals of this class are linked to HBR test indicating which factor among them has influenced variation of HBR. Food quantity: indicates the quantity of a certain type of food consumed by the pregnant women. Medication: an instance of the medication class will be used to ask and the pregnant women about the type of medication taken, food, iron.

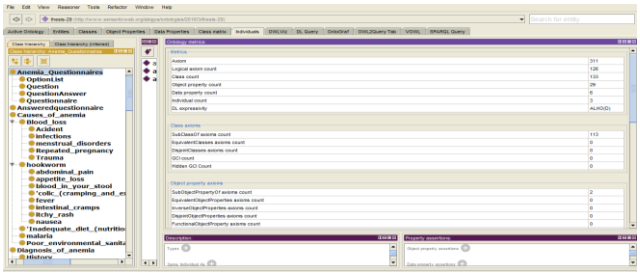


FIGURE3.3: CONCEPT OF ANEMIA ONTOLOGY

3.1. Data Properties

Data properties link individual of class within a domain to its data value while object properties link individuals to individuals or link an individual to an XML schema datatype value or an RDF literal. In this section, we present some data properties that have been used in ontology development. Data properties play fundamental role when describing individual's characteristics and enabling data value to be saved. The data properties used in our model are presented in HasAge, haslocation, hasGender, hasabortion, hasbloodloss, hascontraceptive, hasbednet, hasinfected with malaria and HasTelephone are Data type property used for providing additional information to individual of patient (pregnant women) class in order to distinguish them. Each data type has its own data format. Age and phone number should have integer value, while address and gender have a string value and contraceptive, blood loss and haslatrine and hastea or hascoffee is Boolean. In the Figure 3.9 below, we have described some data properties of individual called almaz.

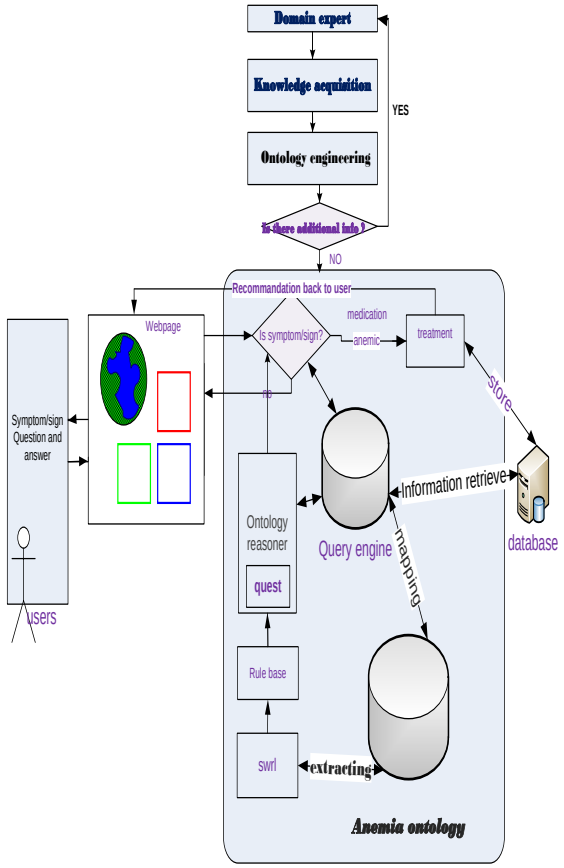


FIGURE3.4: ARCHITECTURE OF MY MODEL

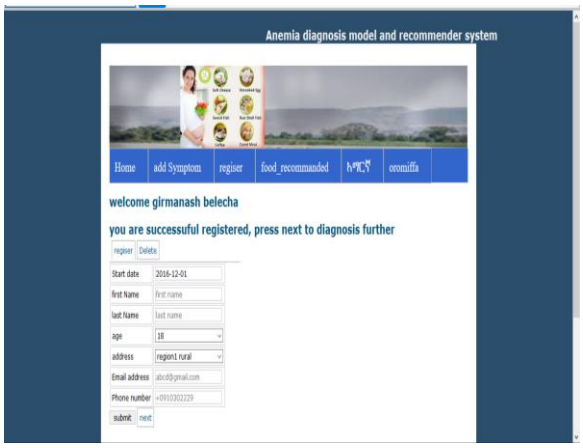


FIGURE3.5: REGISTRATION FORM IN THREE LANGUAGE USERS

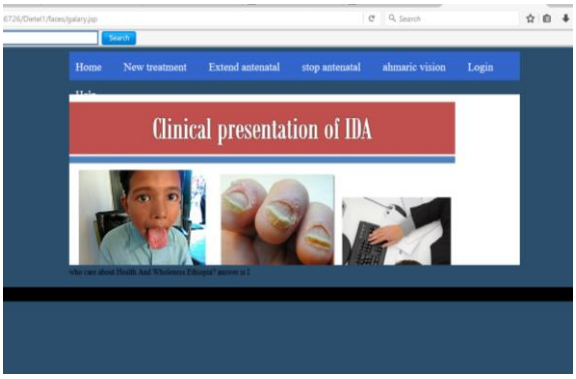


Figure3.6: model show symptom of anemia

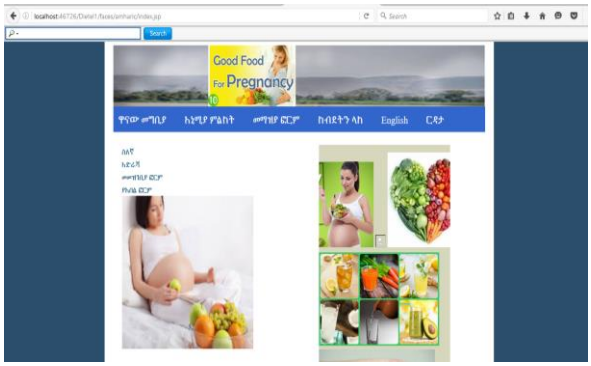


Figure3. 7: main home for Amharic user

Figure 2.7 shows that homepage for Amharic user and the model enables pregnant women's or doctors' diagnosis in Amharic language. The module interface is built in Amharic and English language. Figure 4.5 starts with sociodemographic diagnosis. The model enables pregnant women to fill the forms and when submit button is pressed, the model starts diagnosis and displays result soon to the user.

2.7. Model Evaluation and Testing

The developed anemia ontology is tested and evaluated to check whether the objectives of the research are achieved or not. The evaluation and testing issue of the ontology is summarized by the question "does ODAPW acceptable, correct and accurate recommended service to diagnose anemia?" We use a visual interaction evaluation technique for the purpose. The developed prototype of ODAPW is tested to ensure the performance of the model in meeting towards establishing objectives. Testing is an important step to evaluate the performance of our model. The evaluation process is more concerned with model user acceptance validations of the prototype. User acceptance efforts are concerned with issues impacting how well the model addresses the need of the user. For the user acceptance testing, we prepare domain expert in the area of the paper and model testing, we were testing how the accuracy of the model. After we test our prototypes we evaluate our model. The domain expert evaluators interact with the model by using appropriate cases. Then they evaluated the model by using closed and open ended questionnaires. To assess human factors visual interaction together with questionnaire methods is used.

2.8. Model Testing and Evaluation

Different aspects of ontology can be evaluated: The satisfaction of users when testing the ontology. For example, does the ontology help them to process more rapidly in their task? The completeness of the domain representation:

Is there any ambiguous concept/relation, missing concept/relation or superfluous concept/relation? The correctness of the knowledge base and its inference engine. Does the inference engine produce full knowledge? Of the domain representation: is there once a model knowledge based system is developed, it should be tested and evaluated to make sure that the acceptance and the performance of the model is precise. Testing and evaluation of the model is the final step that assists the knowledge engineer to measure that whether the model is met the proposed objective or not. More importantly, evaluation is carried out to determine users' acceptance and applicability of the model knowledge based system in the domain area.

2.9. User Acceptance Evaluation

In this study, user's evaluation is the process of ensuring that whether the model satisfies the requirements of its end-users and helps to evaluate the performance of the model from the user's perspective. This allows the end-users to make evaluations and comments while interacting with the system(Roussey, 2005). For the purpose of user's evaluation process, five domain experts from

afelegate hospital selected as model evaluators. These experts are selected purposefully from the pregnant woman pathologist. Before starting the evaluation, the researcher explained the objective of the developed model and how the model interacts with the users. This explanation helped the experts to avoid the variation of awareness among them about ontology-based decision support model to diagnosis anemia in pregnant women. After the discussion of the model, to evaluate the user acceptance of ontology-based decision support model to diagnosis anemia in pregnant women, the questionnaires were distributed. Using questionnaires distributed domain experts' give feedback towards developing a model and their feedback is used for analysis. The type of questionnaires distributed for feedback collection for the evaluators is close ended and open ended questionnaires. The evaluators assess the accuracy of the model by using simplicity to use a model and interact with the model, efficiency in time, accuracy of decision made by the model, does the model has adequate knowledge to diagnose anemia disease, the ability of the model in the diagnosis and treatment of anemia disease, easy to learn/adopt new knowledge, and the importance of the model in the domain area. These evaluation formats are customized from (Solomon, 2013; S. Staab, Studer, & 1998). The questionnaires used to test the performance of the model by domain experts are. For the ease of analyzing the relative performance of the model based on user's evaluation, all closed ended questions are answered as excellent, very good, and good, design process refers to closed die forging process. Historically die design was based on trial and error and experiences of tool makers were appreciably utilized.

III. CONCLUSION

There is an enormous gap in fusing ontology-development support in health care services in the developing countries like Ethiopia. Such dodges are the result of many dominant, influential factors like quality of health care, shortage of the distribution of physicians per patient, dearth of experts in the area, the growth rate of population, the unwillingness of doctors rendering their professional services, scarcity of health care units, overcrowded patients, shortage of budget with the public sectors and the high cost of private health care. As a consequence, there is an urgent need of a suitable model that can fill the gap of the existing health care problem in the developing countries especially in Ethiopia.

A significant advantage of our approach is the utilization of the widely-supported mechanism for ontology-development supper for anemic pregnant women. The approach is an apposite tool for the knowledge retrieval and ontology management in healthcare units to share expert knowledge of physicians in medical field to enhance the diagnosis of anemia within semantic web. This research is focused in one particular direction, namely using the ontology for building, maintaining and querying a distributed enterprise knowledge map. In order for our approach to work in a health organization, we assume that it has the accessibility to the Internet.

The model can be used by any user at ease as the model is built on the web within continues button leading forward menu for its usage.

Using our ontology-based GUI generator, a domain expert can browse the data schema using his/her own professional expertise, choose attribute fields according to needs, and make a highly-customized GUI for end users, without having any exposure to programming skills. Our approach has established a systematic methodology for domain experts to specify the business needs in an unambiguous manner. Anemia diseases are curable diseases that require long term supervision and treatments by medical professionals. The most common associated factors of anemia are malaria, hookworm, nutrition, and latrine. With information and communication technology, many applications have been implemented to facilitate different clinical decision making process. With our model, personalized healthcare systems are in place to enable patients with anemia diseases to acquire continuous and long-term medical services at home. This improves health care delivery for medical services can be accessed at pace. Today high prevalence of anemia diseases poses technological challenges to existing personalized healthcare systems. So, our model is appropriate for current circumstances to diagnose the clinical symptoms of anemia and its associated factors such as hookworm, malaria, dietary, mother history and health status of pregnant women and recommendation plan to pregnant women. In addition, the model helps the healthcare units and pregnant women to diagnose anemia where there is an internet facility.

Recommendation and Future Work

- The study achieves its objectives by providing the treatment and diagnosis service to the pregnant women in suffering from anemia. Based on the findings of the study, the following recommendations are suggested for further study on the applicability of ontology-development for anemic pregnant women.

- The recommendation can inspire interested researchers to investigate further implementation of the prototype of ontology-development for anemic pregnant women in related health domain.
- The scope of the ontology-development for anemic pregnant women can be extended to include other affecting factors such as aplastic anemia, sickle cell anemia and Hemolytic anemia.
- Ontology-development for anemic pregnant women can handle any domain specific problem, if there is a perfect knowledge. But, most of the time, the information we gain from the patient may not satisfy the conditions of the given rule. Therefore, it needs a model that can deliver a better solution based on the few respondents' response applied on case-based ontology techniques.
- ODAPW can be used for the purpose of self-treatment among the patients using three different specific local languages. But the speakers and patients can express their feelings using their own languages. Therefore, a user interface of the model should be designed to enable the users to communicate using their own language with the model
- The ODAPW is developed for PC users. It does not include mobile user. Therefore, it is better to include mobile user too.
- In our study, we apply the rule based model which solves problems from the scratch, while case based model uses pre-stored situations to deal with similar new instances. Therefore, the integration of rule based ontology reasoning model with case based ontology reasoning model would solve the limitation when representing knowledge in the form of "if then" clause.
- In our study, we exclude non-pregnant women, child and pregnant women who are severely sick because of some medical conditions like diabetes, renal or cardio- respiratory diseases, HIV/AIDS and anemia hypertension, for which follow up is required. Interested researchers can expand the scope of the model by applying the same approach on such other chronic diseases.
- Further development of a model for end user interface using the ontology-based decision supporting a model in semantic web technology can be extended using tool *sesame*.
- Our model can also be extended by using fuzzy set with which, in our view, a better result may be achieved.

REFERENCE

ALFONSE, M., M., M., AREF, M. M., M., B., & salem, A. B. An Ontology-Based Cancer Diseases Diagnostic Methodology.

Alfonse, M., Mostafa , M., Aref, M. M., Badeeh, M., & Salem, A. B. (2001). An ontology-Based Cancer Diseases Diagnostic Methodology. *Recent Advances in Information Science*.

Alharbi, R. F., Berri, J., & Masri, S. (2015). *Ontology based clinical decision support system for diabetes diagnostic*. Paper presented at the Science and Information Conference, London, UK.

Boose, J. H., & Gaines, B. R. (1989). Knowledge Acquisition for Knowledge-Based Systems: Notes on the State-of-the-Art. *1989 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands., 4*, 377-394

Chen, R. C., & Bau, C. T. (2013). An ontological approach for guideline-based decision support system. *International Journal of Computer, Consumer and Control (IJ3C), 2*, 3.

Ejeta, E., Alemnew, B., Fikadu, A., Fikadu, M., Tesfaye, L., & Birhanu, T. (2014). Prevalence of Anaemia in Pregnant Womens and Associated Risk Factors in Western Ethiopia *Food Science and Quality Management, 31*(), 11.

Emberey, C. L., Milton,N.R,, & Vermeulen, B. (2007). Application of Knowledge Engineering Methodologies to Support Engineering Design Application Development in Aerospace. *Published by the American Institute of Aeronautics and Astronautics, Inc., with permission.*(7th AIAA Aviation Technology, Integration and Operations Conference (ATIO) 18 - 20 September 2007, Belfast, Northern Ireland).

Josep Blat, J., Ibáñez, T., & Navarrete. Introduction to ontologies and tools; some examples.

Pinto , H. S., & Martins, J. P. (2000). Reusing Ontologies. *AAAI Technical Report SS-00-03. Compilation, AAAI Technical Report SS-00-03*.

Potter, S. A Survey of Knowledge Acquisition from Natural Language. *TMA of Knowledge Acquisition from Natural Language*.

Roussey, C. (2005). Guidelines to build ontology a Bibliographic study. *article*.

Solomon, G. (2013). A Self Learning Knowledge Based System For Diagnosis and Treatment of Diabetes.

Staab, S., Studer, R., & (1998). handbook on ontology. *International Handbooks on Information Systems, 2*, 161-197.

Staab, S., studer,R. (2001). Knowledge Processes and Ontologies. *IEEE INTELLIGENT SYSTEMS*.

Studer, R., Abecker, A., & Grimm, S. (2007). *Semantic Web Services Concepts, Technologies, and Applications* (R. Studer, Grimm ,S·, & A. (Eds.)A. Eds. Vol. part I and part II). German: springer.

TSEHAYU, B. T. (2009). *Determinants of anemia in pregnant women with the emphasis on intestinal helminthic infection at bushulo health cnter in souther Ethiopia.* (Masters of Science in medical parasitology.), Addis ababa university school of graduate studies.

wright, R. G. (1992). KNOWLEDGE AND INFORMATION INTERFACE STANDARD for TEST AND DIAGNOSIS APPLICATIONS OF KNOWLEDGE-BASED SYSTEMS.