A Data Mining Based Ontology Approach for Predicting the Research Ideas using Past Research in the Wildlife Sector of Sri Lanka

Premisha Premananthan
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
ppremisha@std.appsc.sab.ac.lk

Kumara BTGS
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
kumara@appsc.sab.ac.lk

Enoka P Kudavidanage
Department of Natural Resources
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
enoka@appsc.sab.ac.lk

Banujan Kuhaneswaran
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
bhakuha@appsc.sab.ac.lk

***Abstract*— Sri Lanka, being a global biodiversity hotspot, places great emphasis on biodiversity from an ecological perspective, socio-economic, and cultural factors. However, the wildlife of Sri Lanka is critically threatened due to several factors, mainly human activities and needs dire conservation measures. Lack of knowledge and technical support also hinder wildlife management activities. Findings of wildlife research studies could be incorporated into data-driven conservation and management decisions but the current contribution is not satisfactory. This research shows a novel data mining approach for finding hidden keywords and automatic labeling for past research work in this domain. We used Latent Dirichlet Allocation (LDA) algorithms to model topics and identify the major keywords also developed an ontology model to represent the relationships between each keyword. These approaches are also useful for future research proposals, for recognizing research holes and can classify the subjects related to a publication by non-professional related fields. The experiment results demonstrate the validity and efficiency of the proposed method.**

***Keywords*— wildlife, LDA, ontology, topic modelling**

I. Introduction

Wildlife is critical for the sustenance of life on earth. Biodiversity conservation is crucial to preserving a stable global ecological balance.

Sri Lanka is a global biodiversity hotspot consisting of a large variety of fauna and flora. It is one of the main sources of income generation through tourism and other means. The diversity of ecosystems is primarily due to its topographical and climatic heterogeneity, as well as its coastal effect [1]. This rich biodiversity is threatened due to unplanned land use, pollution, overexploitation, etc.

Data from wildlife research can contribute to a large extent is proper conservation and management. However, there is a gap between research and application. Most of the existing research work is not converted into applications while there are many data gaps. Limited numbers of researchers are focussing on the actual research needs from conservation. The selection of research topics is often not compatible with the actual research needs due to multiple reasons. This is a disheartening scenario as there are plenty of opportunities for such work. Inadequate knowledge of the existing research and their applicability, inadequate use of technology, and inability to locate some research are some of the contributing factors. Other than the research published in a known journal, some past research information available online cannot be found properly because they belong to conventional archives, unfortunately.

Increasing public awareness on the values of wildlife and the consequences of losing this heritage can assist conservation to a large extent. To achieve this, we have to simplify the gap between the public and the accessibility to information on wildlife. Technology can play a major role in filling the gap between them.

Mostly wildlife studies aimed to understand species diversity, behaviour, and habitat use, and ecology, the role of wildlife in disease transmission, species conservation, population management, and methods to control threats to diversity.

In our study, we focus on analysing past research papers using data mining techniques to give potential research ideas to the future. To fill the data demands for conservation our solution focuses mainly on semi-automating the finding of research gaps via abstract analysis. Finally, the model includes the most commonly used keywords and question top. This will be an important milestone for researches as well as wildlife activists to give tips on recent problems that need a solution urgently.

From a technological perspective, there was prior work [2] [3] that has shown hierarchical relationship-based latent Dirichlet allocation (hrLDA), a data-driven model of hierarchical topics to acquire terminology ontology from a large number of amalgamate documents. Unlike traditional topic models, hrLDA relies on noun phrases instead of unigrams, deals with syntax and text structures, and enriches topic hierarchies with topic relations. Through a series of experiments, we are demonstrating hrLDA's superiority over established topic models, especially for hierarchy building.

So we have to vary past research techniques to find with our final solution. Some trending techniques are used here to improve the outputs. Our research mainly focus to resolve the inadequate application of wildlife research and technologies in the decision-making process. This paper is organized as follows. In Section II, we describe the core theories used by the proposed methodology, and we discuss the results of the study in Section III. In Section IV, we discuss the conclusion of our experiment and suggest areas for future study at the end.

II. Methodology

We used a semi-automated methodology which shows in Fig 1. This methodology developed using Latent Dirichlet Allocation (LDA) and Ontology in this study. The text data of the defined domain were collected and pre-processed for the input to LDA algorithms then compared with the ontology graph to the final output. The steps of our methodology are defined below.

A. *Data Collection*

We collected information about past wildlife researches in Sri Lanka from 2006 to 2019, with the aid of the Department of Natural Resources, Sabaragamuwa University of Sri Lanka, and an extreme literature survey. After that, we accessed full research papers of selected papers from each domain. We've selectively applied the title and abstract data to the CSV file from those research papers.

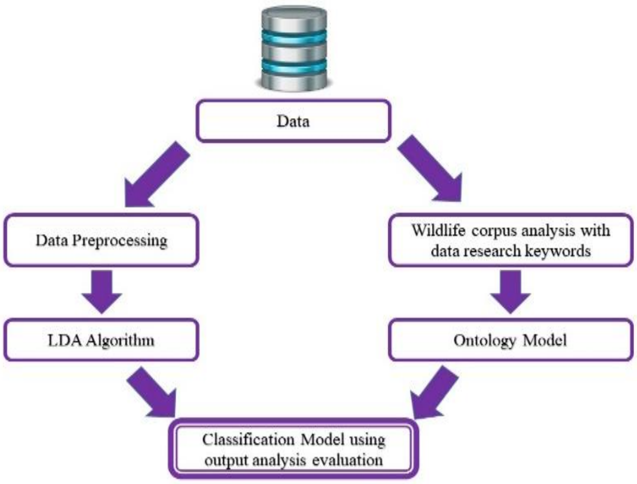

Fig. 1. Methodological framework

B. *Data Pre-processing*

Data pre-processing is so important because if our data set contained mistakes, redundancies, missing values, and inconsistencies that all compromised the integrity of the set, we need to fix all those issues for a more accurate outcome [4]. We performed the following steps:

- Tokenization: Divide the text into sentences, and the sentences into words. Lower case the words and smooth punctuation
- Stop word removal: Delete words that have fewer than 3 letters. All stop words are removed.
- Lemmatizing: Words in the third person are shifted to first-person and verbs shifted to present from past and future tenses.
- Words are stemmed — words are reduced to their root form

C. *Topic Modelling-LDA*

LDA helped adapt the textual data into a format that could act as an input to the LDA model for training. We began by converting the documents to a simple representation of the vectors as a group of words called Bag of Words (BOW) [6]. First, we translated a list of titles into vector lists, all with vocabulary-capable lengths.

The Fig. 2 has shown the topic modelling which is one of the unsupervised methods. In other words, it is a text mining strategy in which the subjects or themes of documents can be derived from a broader set of documents [5]. LDA, one of the most popular modelling techniques, is a probabilistic model of a corpus-based on Bayesian models. This is often considered a probabilistic extension of Latent Semantic Analysis (LSA). The LDA's basic idea is that each document has a word distribution that can be defined as.

D. *Ontology Modelling*

Ontologies contain features such as general vocabulary, reusability, machine-readable content, as well as ordering and structuring information for the Semantic Web application, enabling agent interaction, and semantic searching [7]. Automated learning is the problem in ontology engineering, such as the lack of a fully automated approach to shape ontology using machine learning techniques from a text corpus or dataset of various topics.

The ontology model was finalized using protégé tools, which is the most popular tool of ontology visualization [8]. The Protégé 5.5.0 tool is being applied for further development in various disciplines for a better understanding of knowledge with the aid of domain professionals in the wildlife.

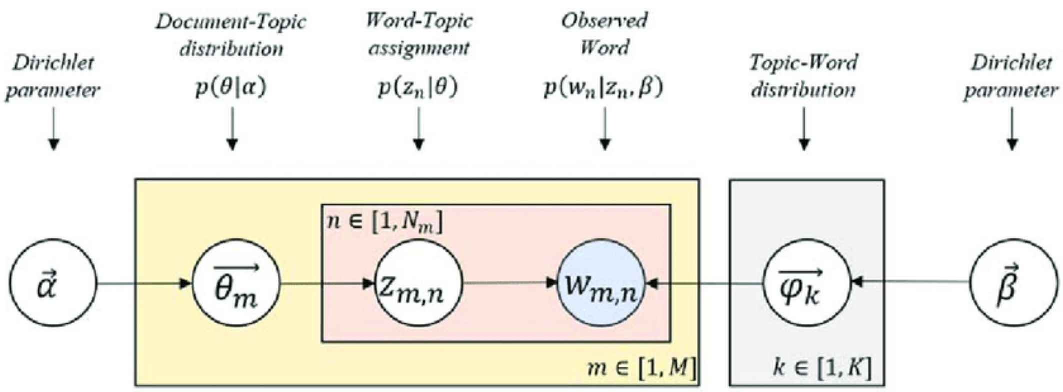

Fig. 2. Graphical model for LDA

Source: Adapted from [5]

E. Comparison

Our interactive, web-based visualization framework, LDAvis, has two key functionalities that allow users to understand the topic-term relationships within a fitted LDA model, as well as several additional features that provide additional perspectives on the model [9]. First and foremost, LDAvis allows one to pick a topic to report the words most applicable to the subject. We compared the total term frequency to the approximate term frequency for finding the keywords that appear and are most significant.

F. Evaluation

In our research, we used output analysis method which used to assess the outcomes of the research in relation to its objectives. We have defined a new approach to automatic ontology learning, and this method has applied the LDA model to generate topics, and the progress of learned ontology does not need the seed of ontology, but only the document corpus [10].

III. Results

The results of this study were represented using abstract past research which serves as an input in Sri Lanka. We used python language for LDA implementation. The text used as input is interpreted and tokenized with the result that input nouns, adjectives and verbs are compiled. In addition, it removes all the stop words in the research papers.

The tokenized and pruned text is then subjected to the LDA modelling algorithm. That gave production as word sets that could collection contain words that are linked to each other. Such collections of words are classified as various subjects. The LDA model approach is used to arrange, synthesize broad corpus, and to retrieve subjects and words.

Fig. 3 and Fig. 4 are the final visualizations of the LDA model which shows the overall keyword for each research paper and the essential keyword using the pyLDAvis library in python. This output allowed the detection of hidden keywords from every abstract. To get the output of the pyLDAvis method we used the equation of saliency and relevance to accommodate the keyword distributions.

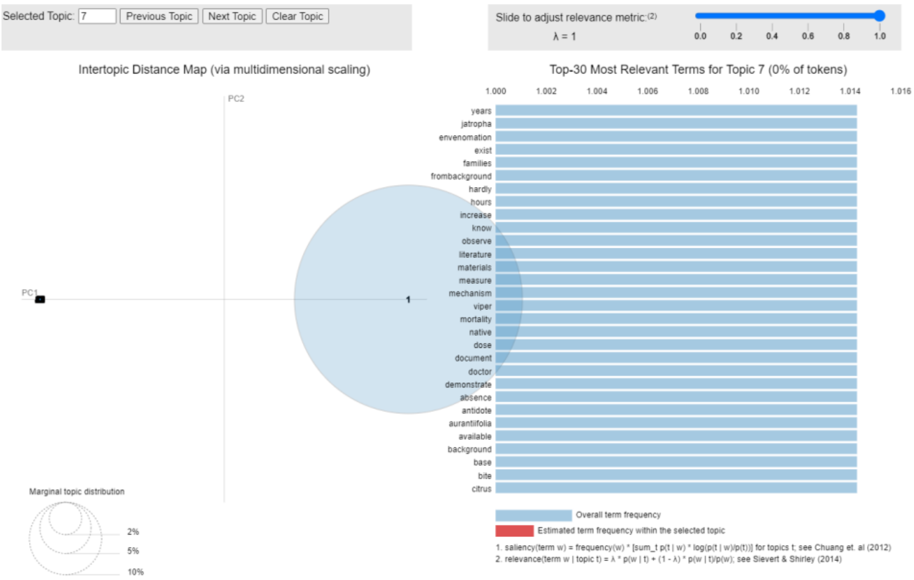

Fig. 3. LDA model for overall keyword

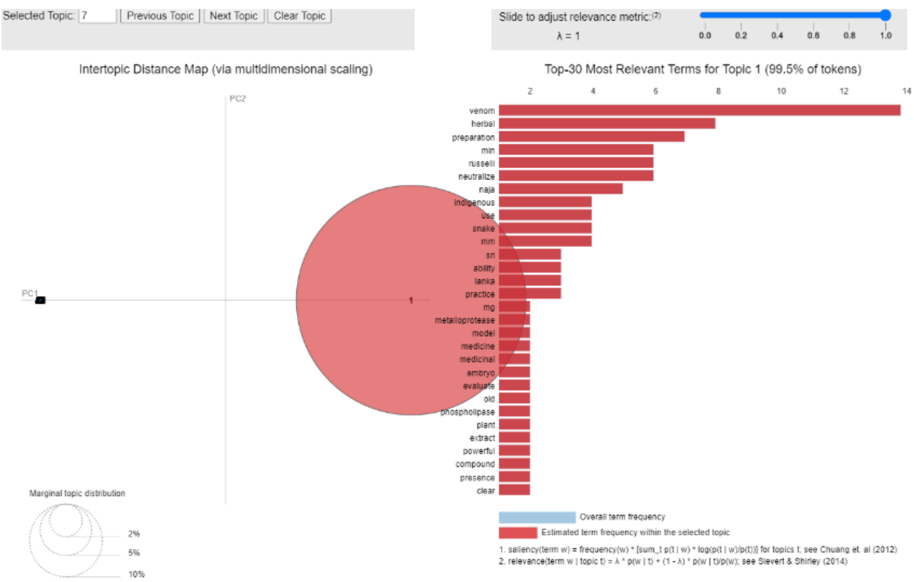

Fig. 4. LDA model for estimated keyword

The intertropical distance map is indicated via multidimensional scaling by our LDA output. In CE literature and inter-topic distance, the top 20 salient keywords.

$$Saliency = frequency \times \left[sum\ p(t|w) \times \log \left(p\left((t|w) / p(t) \right) \right) \right] \quad (1)$$

Where (1), t- Topic, Frequency (w) –frequency of word w, p (t|w) - conditional probability: the likelihood that observed word w was generated by latent topic t, p (t) - the probability of topic t, sum p (t|w) - summation of the probability of observed word w was generated by latent topic t

This formulation (1) defines (in a theoretical context of information Sense) how informative the specific term w, versus a randomly selected word, is for determining the generating subject. For instance, if a word w appears in all topics, observing the word tells us nothing about the topical mixture of the document; thus the word will obtain a score of low distinctiveness. The saliency [11] of a term is defined by the product:

$$Relevance = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w) \quad (2)$$

Where (2), λ –slide to adjust relevant metric, p (w|t) - conditional probability: the likelihood that observed word w was generated by latent topic t, p (w) –the probability of word w [12]

Using this output from LDA we compared the ontology output. Analysed the estimated keywords and their ontology domain formation. The protégé tool used the Sri Lankan wildlife research domain ontology to be developed. The partial view of the final ontology production shown in Fig 5 and Fig 6.

Each research papers' keyword generated by LDA visualization model estimation was analyzed through the ontograph and each paper classification performed. Table I shows the partial final outcome of the study. We manually compared the results from both LDA and terminology ontology. Our study's output evaluation was 92% accuracy to the overall conclusion.

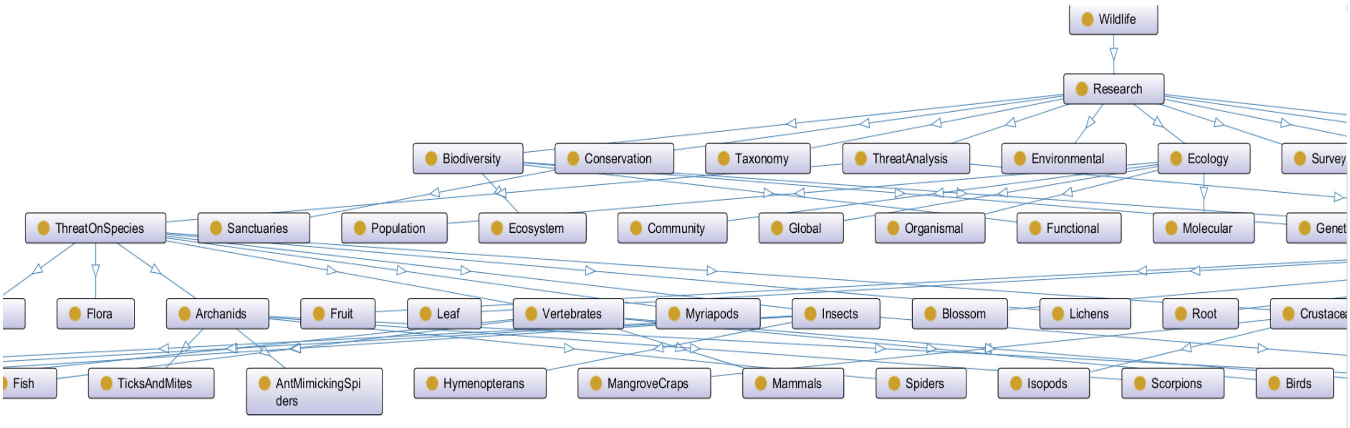

Fig. 5. Ontograph partial view

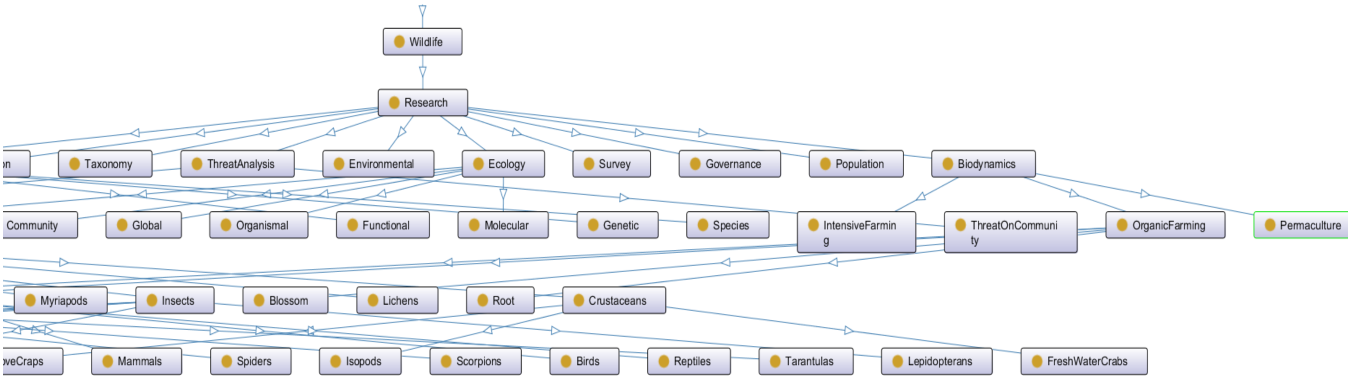

Fig. 6. Ontograph partial view

TABLE I. PARTIAL FINAL OUTPUT

Paper Name	Top 3 Main keywords	Main classes
Characterization of Daboia russelii and Naja naja venom neutralizing ability of an undocumented indigenous medication in Sri Lanka [13]	Venom, Herbal, Preparation	Reptile, Conservation
Marine bacteria and fungi as sources for bioactive compounds: present status and future trends [14]	Marine, bacteria, fungi	Biodiversity, ThreatOnSpecies
Reptile diversity in beraliya mukalana proposed forest reserve, galle district, sri lanka [15]	Forest,reptile, anthropogenic	Environment, Reptile
Changes in soil carbon stocks under different agricultural management practices in North Sri Lanka [16]	Soil,fertilizer, fraction	Ecology, Permaculture

IV. CONCLUSION AND FUTURE WORKS

In this paper, we have suggested a domain-independent and self-learning model which means that the study of ontology in new fields is very exciting therefore can save considerable time and effort in the acquisition of ontology. For past research papers using terminology ontologies, we have established a new approach for automatic classification, and this method applied the LDA model to generate topics, and the progress of learned ontology does not need the seed of ontology, it only requires given document corpus. We generated LDA keywords for selected research abstracts of the past wildlife domain in Sri Lanka. We devise a semi-automated topic labeling for the research papers. The final experiment has proved effective results.

This work reduced the complexity to label the research papers without any domain pre-knowledge. Using this method the hidden keyword and the relations between the keywords also identify to help future research ideas.

In this topic labelling method, there is some inefficient while ontology classification. Because there are several cross path hierarchy moves of keywords identified from LDA. So when we used ontology it collapsed the different path in onto graph. So we will use other classification methods for fully automated our methods.

ACKNOWLEDGEMENT

We acknowledge the Department of Wildlife Conservation of Sri Lanka for the research permit (WL/3/2/60/15) granted to E. P Kudavidanage.

REFERENCES

[1] L.P.Jayatissa, *Present Status of Mangroves in Sri Lanka*. 2012.

[2] X. Zhu, D. Klabjan, and P. N. Bless, "Unsupervised terminological ontology learning based on hierarchical topic modeling," *Proc. - 2017 IEEE Int. Conf. Inf. Reuse Integr. IRI 2017*, vol. 2017-Janua, pp. 32–41, 2017, doi: 10.1109/IRI.2017.18.

[3] S. Chowdhury and J. Zhu, "Towards the ontology development for smart transportation infrastructure planning via topic modeling," *Proc. 36th Int. Symp. Autom. Robot. Constr. ISARC 2019*, no. Isarc, pp. 507–514, 2019, doi: 10.22260/isarc2019/0068.

[4] N. T. R. Editors, *Technologies in Data Science and Communication*. 2019.

[5] J. Lee, J. H. Kang, S. Jun, H. Lim, D. Jang, and S. Park, "Ensemble modeling for sustainable technology transfer," *Sustain.*, vol. 10, no. 7, 2018, doi: 10.3390/su10072278.

[6] M. Rani, A. K. Dhar, and O. P. Vyas, "Semi-automatic terminology ontology learning based on topic modeling," *Eng. Appl. Artif. Intell.*, vol. 63, no. August, pp. 108–125, 2017, doi: 10.1016/j.engappai.2017.05.006.

[7] D. Movshovitz-Attias and W. W. Cohen, "KB-LDA: Jointly learning a knowledge base of hierarchy, relations, and facts," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 1, pp. 1449–1459, 2015, doi: 10.3115/v1/p15-1140.

[8] G. Hussein, A. L. I. Ahmed, L. Kovács, G. Hussein, and A. Ahmed, "ONTOLOGY DOMAIN MODEL FOR E-TUTORING SYSTEM," vol. 5, no. 1, pp. 37–44, 2020.

[9] R. Adhitama, R. Kusumaningrum, and R. Gernowo, "Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, pp. 247–251, 2018, doi: 10.1109/ICICOS.2017.8276370.

[10] Z. Lin, "Terminological ontology learning based on LDA," *2017 4th Int. Conf. Syst. Informatics, ICSAI 2017*, vol. 2018-Janua, no. Icsai, pp. 1598–1603, 2017, doi: 10.1109/ICSAI.2017.8248539.

[11] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," *Proc. Work. Adv. Vis. Interfaces AVI*, no. June, pp. 74–77, 2012, doi: 10.1145/2254556.2254572.

[12] T. R. Frasier, S. D. Petersen, L. Postma, L. Johnson, M. P. Heide-Jørgensen, and S. H. Ferguson, "Bayesian abundance estimation from genetic mark-recapture data when not all sites are sampled: an example with the bowhead whale," *bioRxiv*, vol. 22, p. 549394, 2019, doi: 10.1101/549394.

[13] M. M. Silva, S. S. Seneviratne, D. K. Weerakoon, and C. L. Goonasekara, "Characterization of Daboia russelii and Naja naja venom neutralizing ability of an undocumented indigenous medication in Sri Lanka," *J. Ayurveda Integr. Med.*, vol. 8, no. 1, pp. 20–26, 2017, doi: 10.1016/j.jaim.2016.10.001.

[14] V. Gunathilake., "Marine Bacteria and Fungi As Sources for Bioactive Compounds: Present Status and Future Trends.," *Int. J. Adv. Res.*, vol. 5, no. 9, pp. 610–614, 2017, doi: 10.21474/ijar01/5367.

[15] D. M. S. S. Karunarathna and A. A. T. Amarasinghe, "Reptile diversity in Beraliya Mukalana Proposed Forest Reserve, Galle District, Sri Lanka," *TAPROBANICA J. Asian Biodivers.*, vol. 4, no. 1, p. 20, 2012, doi: 10.4038/tapro.v4i1.4378.

[16] R. R. Ratnayake, T. Kugendren, and N. Gnanavelrajah, "Changes in soil carbon stocks under different agricultural management practices in North Sri Lanka," *J. Natl. Sci. Found. Sri Lanka*, vol. 42, no. 1, pp. 37–44, 2014, doi: 10.4038/jnsfsr.v42i1.6679.