# FCROS: An Approach for Favorable Crop Recommendation Using Ontology Semantics and Selective Knowledge Stack

**Gerard Deepak[1],Anirudh M[2], Akhil S Krishnan[2], Y Pushpanjali[3], SheebaPriyadarshini J[4], Santhanavijayan A[3]**

[1,3,4]Department of Computer Science and Engineering

[1]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

[2]State University of New York, Buffalo, United States of America

Kings College London, United Kingdom

[3]National Institute of Technology , India

[4]CHRIST (Deemed to be University), India

gerard.deepak.cse.nitt@gmail.com



## Abstract

In the present-day, crop recommendation is of utmost importance in methodical agricultural practices. The data available in the domain is of a surplus amount, but the knowledge-centric paradigms are scarce. Thereby, in this age of Web 3.0, Crop recommendation is one of the crucial aspects of the digital transformation of agriculture. As a result, An Ontology-driven and Knowledge centric favorable crop recommendations are of extreme value. This work puts forth an Ontology-based semantically directed, favorable crop recommendation system with a selective Knowledge Stack. The system uses Lin similarity, Wu Palmer Index, Cosine similarity and van Belle and Ahmad Index as semantic measures. Support Vector Classifiers, the Random Forest algorithm and AdaBoost is used for Bagging and Classification of the generated feature pool. The experiments have been conducted for Knowledge Stack using Wikidata that were obtained from various agricultural websites and an accuracy of 96.3% is achieved by the approach.

**Keywords:** AdaBoost, Bagging, Cosine Similarity, Knowledge Stack, Van Belle and Ahmad Index.

## 1 Introduction

Agriculture is a vital part of the development of nations, as it is responsible for the production of various food crops. It has been regarded as the main economic sector of developing countries. The present era also involves the farming of fruits, vegetables, dairy, and poultry. Through the usage of AI systems, which are commonly used in agriculture, farmers are able to improve the quality of their harvests. They are also able to identify and control pests and diseases. In addition, these systems can additionally help in crop recommendation, prediction analysis and Crop Yield Maximization. The increasing number of digital activities in agriculture has led to the emergence of various AI systems in the field of agriculture. These include agricultural robotics, predictive analytics, and soil and crop prediction. Farmers are also using sensors to collect data, which is useful in analyzing and improving the operations of their farms.

The availability of this data is also expected to allow the development of AI systems in the field.The concept of crop recommendation is a central part of precision agriculture. It involves identifying the multiple parameters that can be used to improve crop selection. This process helps in identifying the ideal crop for the farmer [1].Ontology-driven crop recommendation approach is an important factor that is vital while recommending crops, Ontologies are of six data types and the entire process proceeds over a knowledge stack which makes the process knowledge-centric as well.

**Motivation:** The need for a semantically driven ontology-based model for recommending crops is in need of the hour. The data is surplus, and the knowledge is scarce. Hence there is a need for Knowledge centric paradigms. Moreover, the overall structure of web 3.0, semantically driven models for the recommendation of crops considering several parametric factors are required.

**Contribution:** An ontology-driven and knowledge-centric approach has been proposed for crop recommendation. Semantic similarities such as Lin similarity, Wu Palmer, Cosine similarity and van Belle and Ahmad Index have been used to extract terms from the document datasets and six different ontologies to extract the features. A feature pool is generated using the yielded entities and ontologies. Bagging is done using Support Vector Classifiers and Random Forest algorithm along with AdaBoost for classification and class conversion which are featured controlled. An Accuracy of 96.3% and False Discovery Rate of 0.05 has been recorded by the proposed approach.

**Organization:**The work in this paper is structured as stated. Section 2 delineates the current work in this domain and their features. Section 3 discusses the architecture of the approach and the steps involved. Section 4 depicts the implementation and comparison of performance with other chosen baseline approaches. Section 5 presents the conclusion of the work.

## 2 Related Works

Nidhi H Kulkarni et al.,[1], has put forth an ensemble approach in recommending Kharif and Rabi crops on a customized dataset using the Random Forest, Support Vector Machine and Naïve Bayes as the learners. Rohit Kumar Rajak et al.,[2], has developed a crop recommendation paradigm on a soil testing lab dataset using SVM and ANN as learners. G Suresh et al.,[3], have developed a Crop Harvest suggestion framework using the calculation of Help Vector Machine which is used to deliver a crop recommendation system on digital farming grounds. V Sellam et al.,[4], have put forth crop yield prediction model with emphasis on cultivated area, rainfall per year and Index of food prices using regression analysis.

S Pudumalar et al.,[5], has proposed a model for the recommendation of crops using an ensemble framework with max voting technique using CHAID, Naïve Bayes, Random tree and K-Nearest Neighbour as learners to recommend crops for site specific paradigms. M Kalimuthu et al.,[6], have put forth a machine learning model for crop prediction that has been used on selected location and space using Naïve Bayes Classifier. Petteri Nevavuori et al.,[7], have presented a model which utilizes a Convolutional Neural Network for crop yield prediction dependent on NDVI and RGB statistics. Saeed Khaki et al.,[8], have developed a trained Deep NeuralNetwork in combination with Shallow Neural Network and Regression Tree for Crop Yield Prediction. In [8-17] several frameworks relevant to the literature of the proposed Framework have been depicted.

## 3 Proposed System Architecture



The for the proposed crop recommendation system is delineated in Fig. 1. The framework is ontology-driven and knowledge centered framework. The model is dataset driven; however, the user preferences and queries are taken into consideration as a viable option in the last step of the recommendation process. Firstly, the dataset is subject to pre-processing. The Pre-processing step involves lemmatization, stop word removal, tokenization of the sentences and named entity recognition. Further to the dataset pre-processing, individual terms from the category dataset, as well as the document dataset, are extracted. These terms while being informative are not feature rich. In order to ensure that these features become more informative, the features are subjected to ontology alignment of six distinct types.

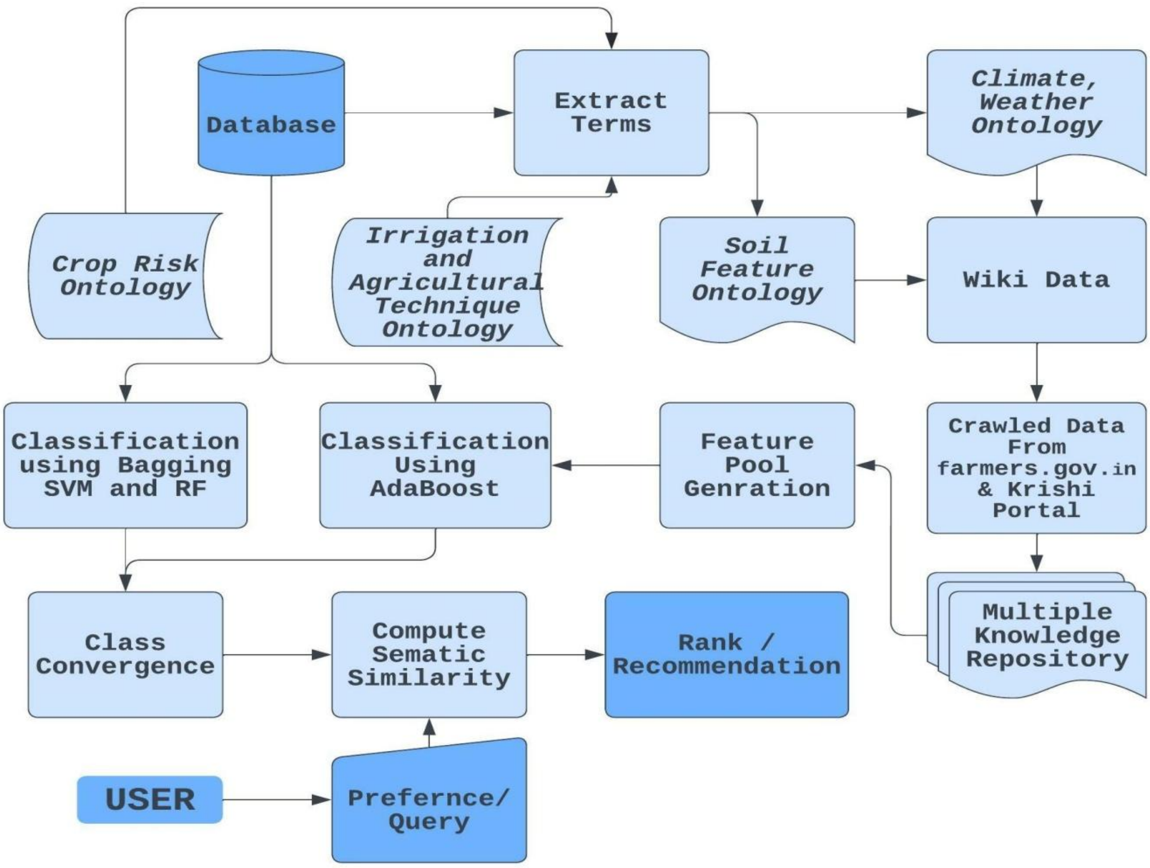




**Fig.1.** System architecture for the proposed FCROS architecture

First, the soil feature ontology is generated and modelled, based on the features of the soil like structure, strength and density. Secondly, climatic zone ontology is taken into consideration which classifies the climatic zone as tropical, temperate, arid etc.; This ontology is modelled comprising several zones and their substantial features. The climate and weather features ontology are typically of how climatic conditions and variations are suitable for ideal crop growth. The agriculture technique ontology comprises several techniques of vegetation and agricultural growth, and is modelled using the WebProtégé ontology creation tool. Similarly, the crop risk ontology and crop irrigation ontology are modelled. The crop risk ontology consists of the attributes which would cause certain degrees of risk in growing crops, and the irrigation ontology consists of the irrigation aspects, suitable and not suitable for crop growth. Most of these ontologies are modelled based on domain expert consultation, agriculturists, and farmer consultations. However, content from geological, geographical textbooks are used as document repositories for automatic generation of ontologies using the Ontocolab framework.

The aforementioned ontologies are subject to computation of semantic similarities with the extracted terms. In order to compute the semantic similarity, Lin similarity and Wu Palmer similarity measures are used with a threshold of 0.5, ensuring a large number of entities are anchored in the initial selection space. The Wu Palmer similarity measure computes the extent of the semantically equivalent data elements in the WordNet database along with the extent of the least common inclusion. The equation for the Wu Palmer similarity is depicted in equation (1)., where LCS is the least common inclusion, and N1 and N2 are the synsets.

$$\text{Wu-Palmer} = 2 \times \frac{\text{dep(LCS(N1,N2))}}{(\text{dep( N1)+dep}(N2))} \qquad (1)$$

Once these ontologies are subject to ontology alignment with terms of the dataset, the entities of the ontologies as well as the dataset selected terms are passed into a knowledge stack. This knowledge stack comprises of Wikidata, the crawled data from farmer.gov.in, the government Krishi portal and the AgriDiksha of the ICAR - Indian Agricultural Research Institute. All these are knowledge repositories pertaining to agriculture, crop growth, etc., from the Indian perspective of farming conditions. The ontologies and the terms are passed via the Wikidata API to enrich the terms, harvesting the hierarchical data. However, the crawls from the farmer.gov.in, the government Krishi portal and the AgriDiksha are semantically related by computation of Shannon's entropy, and are further filtered into the ontologies that are aggregated by computing the cosine similarity, with a threshold of 0.5. Cosine Similarity is quantified by the degree of similarity between a set of documents irrespective of the size. The equation for cosine similarity is depicted in equation (2), where G and F are the two vectors.

$$\cos(\theta) = \frac{\mathbf{G} \cdot \mathbf{F}}{\parallel \mathbf{G} \parallel \parallel \mathbf{F} \parallel} = \frac{\sum_{i=1}^{n} G_i F_i}{\sqrt{\sum_{i=1}^{n} G_i^2}\sqrt{\sum_{i=1}^{n} F_i^2}} \qquad (2)$$

All the entities yielded from the knowledge stack, Wikidata and extracted terms are modelled into a feature pool. The features from the pool are used to classify the dataset using two distinct classifiers simultaneously. The classification takes place using bagging and Adaboost classifiers. The bagging is an ensemble model carried out by using Support Vector Classifier and Random Forest. Bagging is a process of aggregating methods of learning to reduce the divergence in a dataset full of noise. The Support Vector Machine is a type of an algorithm which is supervised to perform classifying tasks and regression, while the random forest classifier incorporates decision trees on the different sub-samples of the provided dataset.

The Adaboost classification begins by training a classifier on the first dataset, followed by training additional replicas of the classifier on the same dataset, with the weights of cases that were mistakenly classified being changed such that later classifiers would concentrate more on challenging situations. This simultaneous classification is done to improve classification and improve heterogeneity of results. The output of the classifiers is feature controlled; this ensures relevance compared to automatic feature selection driven deep learning classifiers. The class convergence is apprehended using the semantic similarity computation of the classes discovered through classification, through cosine similarity with a threshold of 0.7. If the semantic similarity is relevant by a step deviation of 0.25, then the classes and instances under the classes are considered equivalent. Furthermore, the classes considered equivalent are considered for semantic similarity computation with that of the pre-processed user queries or preferences. In this instance the semantic similarity is enumerated using the cosine similarity and the Van Belle and Ahmad Index. The equation for the Van Belle and Ahmad Index is depicted in equation (3).

$$2 * \sum_i \, [p_i q_i/(p_i + q_i)] \qquad (3)$$

The step deviation for cosine similarity is set as 0.75, and 0.25 for the Van Belle and Ahmed Index. Finally, all the entities which are semantically similar are grouped in the rising order of similarity, and the favourable crops are recommended to the user.

## 4 Implementation and Performance Evaluation

The proposed crop recommendation system was implemented with the Python 3.9 programming language, on the Google Collab development environment. An intel core i7 processor and 32 gb of ram were used, along with 8 gb of graphics memory. The ontologies were automatically generated using the OntoColab tool and the webprotege. An integrated dataset comprising of four standard datasets are made use of. Firstly, the Kaggle's Crop Recommendation dataset was used. The second dataset used was the crop recommendation dataset by analyst-2 (http://analyst-2.ai/) tool from the inspirent GmbH. The third dataset used were the land use and cropping recommendations from 1937-1938, provided by the Iowa State University, and the fourth dataset used was the Agro – Meteorological data of Indian state Tamil Nadu. The datasets aforementioned were combined into a single large dataset by annotating them and categorizing them at a single common point.TheAlgorithm for the FCROS system presented is depicted in Algorithm 1.

**Algorithm 1: Algorithm for the FCROS crop recommendation model**

**Input**: User queries, crops database, Aggregated Data Repositories
**Output:** Crop Recommendations based on favourable conditions

```
Pre-process queries, web data and extract terms from ds
ConductOntology Alignment for the extracted features
Model Soil Feature Ontology and Climate Ontology
Model agricultural techniques ontology using WebProtégé
Model Irrigation and Crop risk ontology
while (thresh sim > 0.5) do
    Compute Semantic Similarity of ontologies with extracted terms
end
Form Knowledge stack with aggregation of knowledge bases
Pass entities from ontology and ds terms to the knowledge stack
for (thresh sim > 0.5) in crawls of gov webs and Krishi portal do
    shannon's measure and aggreg ontology using cosine similarity
end
Create Knowledge pool for the entities of knowledge stack
Classify features using distinct classifiers, bagging and Adaboost
for (thresh sim > 0.7) do
    Apprehend class convergence using cosine similarity
end
for equivalent classes and pre–processed terms do
    Calculate cosine semantic similarity (where step deviation = 0.75)
    Calculate Van Belle and Ahmed sim (where step deviation = 0.25)
end
Recommend semantically similar results in increasing order i.e., the favourable crops to user
```

The performance of the proposed FCROS framework, which is a favorable Favorable Crop Recommendation Using Ontology Semantics and Selective Knowledge Stack, is evaluated using Recall, Precision, F-measure percentages, Accuracy and the false discovery rate (FDR) as potential metrics. Precisely, Recall, Accuracy, F-measure percentages indicate and quantify the relevance of the results which are furnished by the framework. The false discovery rate i.e., the FDR quantifies the number of false positives, generated by the framework. Toquantify the performance of FCROS by comparing it with potential models, the CRSMLT [1], CRSET [2] and ECRSDF [3] frameworks were used, illustrated in Table 1.

**Table 1.**Performance of the FCROS approach with baseline models

| Model | Average Precision % | Average Recall % | Average Accuracy % | Average F-Measure % | FDR |
|---|---|---|---|---|---|
| CRSET [1] | 90.79 | 92.63 | 91.71 | 91.7 | 0.1 |
| CRSMLT [2] | 88.22 | 90.45 | 89.34 | 89.34 | 0.12 |
| ECRSDF [3] | 91.18 | 93.07 | 92.12 | 92.12 | 0.09 |
| **Proposed FCROS** | **95.39** | **97.21** | **96.3** | **96.29** | **0.05** |

The FCROS with an average accuracy of 96.3%, an Average F-Measure of 96.29%, Precision and Recall of 95.39% & 97.21% with FDR of 0.05, has yielded the highest performance metrics pertaining to average Precision, Recall, Accuracy, F-measure and the lowest value of FDR, which was evaluated for 3842 queries whose ground truth has been crawled, evaluated, and also collected from 984 participants who were agriculture, horticulture, and botany students.The reason why the proposed FCROS has the highest performance metrics is mainly due to the fact that it is knowledge-centric, semantically inclined, and auxiliary knowledge is fed into the model by means of distinct ontologies, namely the climate ontology, the agricultural techniques ontology, tropical, temperate and other zone ontology, soil feature ontology, crop risk ontology, and irrigation features ontology. So, several features are incorporated into the model by means of an ontology which is semantically aligned using the Wu Palmer similarity measure.

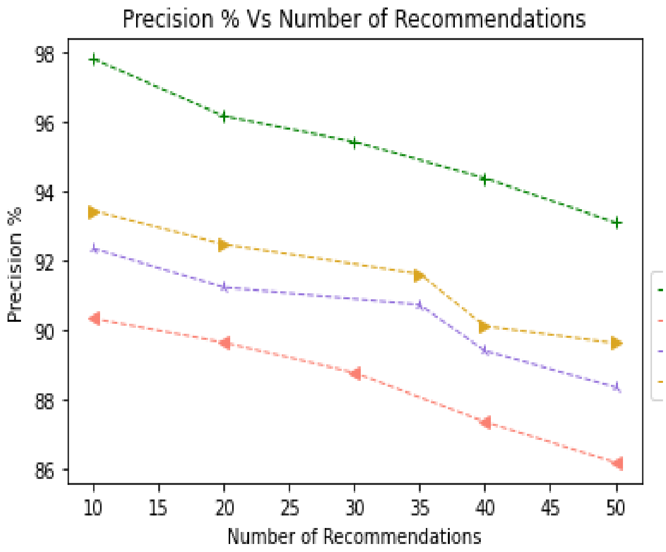


**Fig. 2**. Precision % vs Number of Recommendations

Apart from this, the knowledge stack is prepared by means of using Wikipedia, Krishi Portal, farmer.gov.in portal, which houses heterogeneous bar entities, instances, and entities as auxiliary knowledge which is provided and subsequently yielded as auxiliary knowledge or background knowledge into the model. Apart from this, two distinct machine learning strong classifiers, namely the Badging and AdaBoost classifiers are used. Bagging, already being a strong ensemble model is used for classification by means of controlled feature selection. AdaBoost also incorporates a controlled feature selection scheme for the classification of datasets. Apart from this, there is the inclusion of semantic similarity schemes with varied thresholds ensures the strong relevance of computation mechanisms. The amalgamation of knowledge from several heterogeneous instances of knowledge stack and the several phases of ontologies and incorporation of two distinct heterogeneous classifiers with heterogeneous semantics similarity measures with various thresholds also improve the results.

The CRSET[1] model, is a grouped model using Random Forest, Naive Bayes based classifier, and linear SVM, uses soil-specific features alone and does not use any other alternative features. Apart from this, the CRSET[1] model, which is an ensemble model which uses selective features with strong levels of classifiers, does not augment auxiliary knowledge. Henceforth, the CRSET[1] model also lacks to a large extent. The CRSMLT[2], is a crop recommendations Scheme for Crop Yield maximization, using machine learning algorithms. It takes only one specific vantage point, which includes only the soil databases, from where the soil features alone are taken. Having the scheme of majority voting amalgamated with SVM and neural networks with a strong classification scheme, but only selecting soil features makes the feature selection quite sparse. As a result, the CRSMLT[2] doesn't perform as expected, and moreover, it is only dependent upon the ensemble machine learning schemes without depending upon relevance computation measures. The ECRSDF[3] model uses the help vector machine concept by controlling dirt quality. So, the hyper help vector machine performs as a strong classifier, and the dirt quality factor ensures that there is some kind of selective preferential incorporation of quality knowledge into the model. However, it does not match up to the proposed FCROS model, and the ECRSDF[3] model also lags to a large extent.The Fig.2. delineates the percentage of the precision for the proposed FCROS approach over a ranged value of recommendations, contrastedwith the chosen baseline models. The proposed system for crop recommendation proves superior over the other models owing to selective preferential feature inclusion from ontology. This includes heterogeneity as well as specific inferential knowledge augmenting for inferential learning. The knowledge stack enhances and enriches the later augmented knowledge and the hybridization of bagging with AdaBoost classifiers, semantic similarity and varied heterogeneous semantic measures with differential thresholds.

## 5 Conclusion

An Ontology-driven and Knowledge-centric Crop recommendation system is developed that utilizes an ontology imbued approach with six different ontologies. The proposed model for crop recommendation fixates on pre-processing the data through four different natural language processing techniques. Semantic techniques such as Lin similarity, Wu Palmer Index, Cosine similarity and, van Belle and Ahmad Index are integrated to obtain the best possible data stream that further enriches the knowledge space. A feature pool is yielded using the entities, and classifying features extracted through Bagging. Bagging is carried out using Support Vector Classifiers and Random Forest Algorithm. AdaBoost is made use ofadditionally for classification and the algorithms facilitatethe class conversion process that is feature controlled. Acustom-made Knowledge Stack is created using

WikiData from Agriculture websites like Krishi portal and AgriDiksha that houses heterogenous instances and entities. The experiments have been conducted on the Knowledge Stack that yields an accuracy of 96.3% anda low FDR.

## References

Kulkarni, N. H., Srinivasan, G. N., Sagar, B. M., & Cauvery, N. K. (2018, December). Improving crop productivity through a crop recommendation system using ensembling technique. In 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) (pp. 114-119). IEEE.

Rajak, R. K., Pawar, A., Pendke, M., Shinde, P., Rathod, S., &Devare, A. (2017). Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology, 4(12), 950-953.

Suresh, G., Kumar, A. S., Lekashri, S., &Manikandan, R. (2021). Efficient crop yield recommendation system using machine learning for digital farming. International Journal of Modern Agriculture, 10(1), 906-914.

Sellam, V., &Poovammal, E. (2016). Prediction of crop yield using regression analysis. Indian Journal of Science and Technology, 9(38), 1-5.

Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T., &Nisha, J. (2017, January). Crop recommendation system for precision agriculture. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 32-36). IEEE.

Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020, August). Crop prediction using machine learning. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 926-932). IEEE.

Nevavuori, P., Narra, N., & Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. Computers and electronics in agriculture, 163, 104859.

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in plant science, 10, 621.

Bandara, P., Weerasooriya, T., Ruchirawya, T., Nanayakkara, W., Dimantha, M., &Pabasara, M. (2020). Crop recommendation system. International Journal of Computer Applications, 975, 8887.

Kumar, A., Sarkar, S., & Pradhan, C. (2019, April). Recommendation system for crop identification and pest control technique in agriculture. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0185-0189). IEEE.

Deepak, G., Kumar, N., Bharadwaj, G. V. S. Y., &Santhanavijayan, A. (2019, December). OntoQuest: an ontological strategy for automatic question generation for e-assessment using static and dynamic knowledge. In 2019 Fifteenth International Conference On Information Processing (ICINPRO) (pp. 1-6). IEEE.

Kaushik, I. S., Deepak, G., &Santhanavijayan, A. (2020). QuantQueryEXP: a novel strategic approach for query expansion based on quantum computing principles. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), 573-584.

Hybridised, K. C. N. OntoKnowNHS: Ontology Driven Knowledge Centric Novel Hybridised Semantic Scheme for Image Recommendation Using Knowledge Graph. Knowledge Graphs and Semantic Web, 138.

Yethindra, D. N., & Deepak, G. (2021, September). A semantic approach for fashion recommendation using logistic regression and ontologies. In 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-6). IEEE.

Deepak, G., &Santhanavijayan, A. (2022). UQSCM-RFD: a query–knowledge interfacing approach for diversified query recommendation in semantic search based on river flow dynamics and dynamic user interaction. Neural Computing and Applications, 34(1), 651-675.

Adithya, V., & Deepak, G. (2021, March). OntoReq: an ontology focused collective knowledge approach for requirement traceability modelling. In European, Asian, Middle Eastern, North African Conference on Management & Information Systems (pp. 358-370). Springer, Cham.

Vishal, K., Deepak, G., &Santhanavijayan, A. (2021). An approach for retrieval of text documents by hybridizing structural topic modeling and pointwise mutual information. In Innovations in Electrical and Electronic Engineering (pp. 969-977). Springer, Singapore.