Smart Fields Innovating Agriculture Food and Crop Recommendation

Arvind Kumar[1], Amarjeet Kumar Singh[2], Bijay Rajak[3], Hrithik Sharma[4],

Mr A Suresh kumar[5] , Vishal Kumar[6] , Jivesh Kumar[7]

[5]Assistant Professor, Department of Computer Science and Engineering, Excel Engineering College, Komarapalayam, Tamil Nadu

[1,2,3,4,6,7]Students Department of Computer Science and Engineering, Excel Engineering College, Komarapalayam, Tamil Nadu

---*****************---

Abstract

Data mining involves the extraction of meaningful information from data, and its applications span various domains such as finance, retail, medicine, and agriculture. Within the realm of agriculture, data mining proves invaluable for analyzing both living (biotic) and non-living (abiotic) factors. In the context of India, agriculture plays a pivotal role in the economy and employment sector. A prevalent issue among Indian farmers is the inadequate selection of crops based on soil requirements, resulting in significant productivity setbacks. Precision agriculture emerges as a solution to this challenge, employing modern farming techniques that leverage research data on soil characteristics, types, and crop yields. Precision agriculture assists farmers in making informed decisions by recommending suitable crops based on site-specific parameters, ultimately mitigating the risk of erroneous crop choices and enhancing overall productivity. This paper addresses the farmers' crop selection predicament by proposing a recommendation system. The system utilizes an ensemble model with a majority voting technique, incorporating Random Tree, CHAID, K-Nearest Neighbor, and Naive Bayes as learners. The goal is to recommend crops with high accuracy and efficiency tailored to specific site parameters.

Keywords

Advanced farming techniques, Crop advice system, Collaborative model, Consensus decision-making method, Random tree algorithm, CHAID algorithm, Proximity-based Neighbor analysis, and Probabilistic Bayesian approach.

INTRODUCTION

India, as one of the oldest agricultural societies, has witnessed significant transformations in its farming practices, especially in response to globalization. The evolution of agriculture in the country has been influenced by various factors, leading to the adoption of new technologies aimed at revitalizing the sector. Precision agriculture has emerged as a promising technique in this context, focusing on site-specific farming practices. This approach offers advantages such as enhanced resource efficiency, improved outputs, and informed decision-making in agricultural activities. Despite the positive impact of precision agriculture, certain challenges persist.

In the current agricultural landscape of India, numerous systems have been developed to provide recommendations for specific farming lands. These systems offer guidance on crops, fertilizers, and farming techniques. Among these, crop recommendation is a crucial aspect of precision agriculture, relying on various parameters for accurate decision-making. Precision agriculture endeavors to identify and address these parameters in a site-specific manner to optimize crop selection. While the site-specific approach has yielded improvements, there is an ongoing need for vigilance in assessing the results generated by such systems.

The accuracy of recommendations is paramount in agriculture, as errors can lead to significant material and capital losses. Ongoing research endeavors aim to develop models for crop prediction that are both accurate and efficient. Ensembling, a technique utilized in various machine learning approaches in this field, is proposed in this paper as a method to enhance efficiency and accuracy. Specifically, the paper advocates for a system employing a voting method to construct a robust and precise model for crop prediction.

LITERATURE SURVEY

The first paper [1] delves into the prerequisites and strategic planning necessary for developing a software model dedicated to precision farming. It extensively examines the fundamentals of precision farming, starting from the basics and progressing towards the creation of a supportive model. This paper introduces a model that applies Precision Agriculture (PA) principles specifically to small, open farms at the individual farmer and crop level. The primary goal is to exert a level of control over variability and deliver direct advisory services to even the smallest farmer on their smallest crop plot, utilizing accessible technologies like SMS and email. Although designed for the agricultural scenario in Kerala State, where the average landholding size is smaller than most of India, the model can be adapted for deployment elsewhere in the country with minimal modifications.

The second paper [2] conducts a comparative study of classification algorithms, evaluating their performance in yield prediction for soya bean crops in precision agriculture. The study encompasses Support Vector Machine, Random Forest, Neural Network, REPTree, Bagging, and Bayes algorithms. The conclusion drawn from the analysis is that bagging stands out as the most effective algorithm for yield prediction, exhibiting the lowest error deviation with a mean absolute error of 18985.7864.

The third paper [3] underscores the importance of crop yield prediction in informing strategic agricultural policy-making at the national level. It introduces the eXtensible Crop Yield Prediction Framework (XCYPF), offering flexibility through the inclusion of various techniques for crop yield prediction. The framework includes a tool that assists in predicting crop yield for various crops, considering both dependent and independent variables.

The fourth paper [4] explores the utilization of agricultural data with data mining and visual data mining techniques. The focus is on reducing high-dimensional agricultural data to a more manageable size to extract valuable knowledge related to yield and input application, such as fertilizers. Self-organizing maps and multi-dimensional scaling techniques (Sammon's mapping) are employed for data reduction, with the conclusion that Self-organizing maps are suitable for large datasets, while Sammon's mapping is preferable for smaller datasets.

The fifth paper [5] underscores the significance of crop selection, considering factors such as production rate, market price, and government policies. The proposed Crop Selection Method (CSM) aims to improve the net yield rate of crops by addressing the crop selection problem. The method suggests a series of crops to be selected over a season, considering various factors like weather, soil type, water density, and crop type. The accuracy of CSM relies on predicting the values of influential parameters, highlighting the need for a prediction method with enhanced accuracy and performance.

In the sixth paper [6], data mining techniques are applied to estimate crop yield for cereal crops in major districts of Bangladesh. The methodology involves clustering for creating district clusters and classification using k-NN, Linear Regression, and artificial neural network (ANN) in the rapid miner tool. The accuracy of prediction falls within the range of 90-95%, with a future recommendation for geospatial analysis to further enhance accuracy.

The seventh paper [7] addresses the challenge of selecting classifiers for ensemble learning. The proposed method, Selection by Accuracy and Diversity (SAD), aims to achieve higher accuracy and performance by identifying the dependency between relevant and accurate classifiers using Q statistics. The ensemble is formed by combining classifiers that were not individually chosen, ensuring higher performance and diversity.

The eighth paper [8] introduces various classification methods for classifying liver disease data sets, emphasizing the importance of accuracy in relation to the dataset and learning algorithm. Classification algorithms such as J48, Naive Bayes, ANN, ZeroR, 1BK, and VFI are employed and compared based on their effectiveness, correction rate, accuracy, and computational time. The conclusion is that, except for Naive Bayes, all classifiers demonstrate improved predictive performance, with the multilayer perceptron exhibiting the highest accuracy.

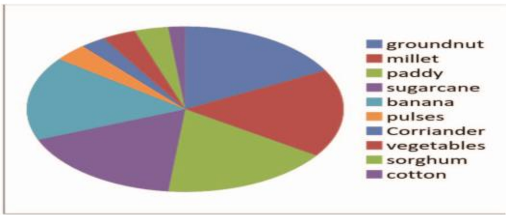
The ninth paper [9] addresses the issue of food insecurity in Egypt by proposing a framework for predicting production and imports for a particular year. Artificial Neural Networks, along with Multi-layer perceptron in WEKA, are employed for prediction, with the aim of visualizing production, import, need, and availability to facilitate decisions on further food imports.

The tenth paper [10] analyzes soil datasets to predict categories, and subsequently, crop yield is identified using Naïve Bayes and k-Nearest Neighbor algorithms. The future work proposed involves creating efficient models using various classification techniques such as support vector machine and principal component analysis.

METHODOLOGY

Dataset Collection

The dataset utilized for this study comprises soil-specific attributes collected from Madurai district, tested at the Soil Testing Lab in Madurai, Tamil Nadu, India. Additionally, we incorporated relevant information from online sources providing general crop data. The crops under consideration in our model encompass millet, groundnut, pulses, cotton, vegetables, banana, paddy, sorghum, sugarcane, and coriander. Figure 1 offers a visual representation of the dataset analysis, presenting the number of instances for each crop available in the training dataset. The attributes taken into account include Depth, Texture, pH, Soil Color, Permeability, Drainage, Water Holding Capacity, and Erosion.



The specified soil parameters play a pivotal role in influencing a crop's capacity to extract essential water and nutrients from the soil. Achieving optimal crop growth is contingent upon the soil providing a conducive environment. Essentially, the soil acts as the anchor for the roots, and its water holding capacity is crucial in determining the crop's ability to absorb nutrients. Nutrients in the soil undergo transformation into ions, a form that plants can readily utilize. The soil's texture, indicating its porosity, governs the movement of air and water, a critical factor in preventing waterlogging and facilitating plant health. Moreover, soil texture influences the soil's capacity to retain essential nutrients. The soil's acidity or alkalinity level (pH) serves as a master variable that significantly impacts the availability of soil nutrients. It further influences the activity of microorganisms within the soil and the level of exchangeable aluminum. The water holding and drainage characteristics of the soil directly affect the penetration of roots. Given these considerations, the aforementioned parameters are crucial in the crop selection process.

Crop Prediction using ensembling technique: Crop prediction employing ensembling techniques involves the use of Committee Methods or Model Combiners, which leverage the collective strength of multiple models to achieve predictions with greater efficiency than any individual model could attain on its own. In our system, we adopt the well-known ensembling technique known as Majority Voting. This technique allows the incorporation of any number of base learners, with a minimum requirement of two base learners. The selection of learners is crucial; they should exhibit competence with each other while also offering complementary capabilities. Increased competition among learners enhances the likelihood of better predictions. However, it is essential for the learners to be complementary, as errors made by one or a few members can be corrected more effectively by the remaining members. Each learner constructs itself into a model, trained using the provided training dataset. When classifying a new sample, each model independently predicts the class. The final class label assigned to the new sample is determined by the majority of learners' predictions. This methodology is implemented using the RapidMiner tool, as illustrated in Figures 2, 3, 4, and 5, showcasing the implemented process in RapidMiner.

Learners Used in the Model:

RANDOM TREE: The Random Tree algorithm, as described in reference [11], shares similarities with traditional decision trees but distinguishes itself by incorporating a unique approach to attribute selection during each split. Unlike a typical decision tree, where all attributes are considered, a Random Tree randomly samples a subset of attributes for each split. This technique is applicable to both nominal and numerical data. In essence, the Random Tree algorithm is akin to established methods like C4.5 or CART, with the key distinction being that, prior to training, it specifically chooses a random subset of attributes for consideration at each node. The parameter controlling the size of this subset is denoted as the subset ratio.

CHAID: In the realm of decision tree techniques, CHAID (Chi-squared Automatic Interaction Detection), as detailed in reference [13], is a distinctive approach founded on adjusted significance testing. While akin to traditional decision trees, CHAID offers several advantages over methods like information gain. One notable advantage lies in its propensity for highly visual and interpretable outcomes. This is achieved through default implementation of multiway splits, contributing to a more comprehensive understanding of the data structure and relationships within the decision tree.

K-NEAREST NEIGHBOR: The K-Nearest Neighbor (K-NN) algorithm, as presented in reference [15], serves dual purposes, catering to both classification and regression tasks. Operating as a straightforward and non-complex algorithm, K-NN retains all available cases in its dataset and makes classifications for new cases by assessing their similarity to existing ones. The classification of a sample set is determined by its "closeness," which is measured through distance metrics such as Euclidean distance or Manhattan distance. In essence, K-NN relies on the proximity of data points to make predictions, leveraging the concept that similar cases in a dataset should exhibit similar outcomes.

NAÏVE BAYES: The Naive Bayes classifier, as referenced in [14], operates as a straightforward probabilistic model. It employs Bayes' theorem, a fundamental concept in Bayesian statistics, with a notable reliance on the naive independence assumption. Naive Bayes serves as a technique for constructing classifier models, assigning class labels to problems by leveraging the principles of Bayesian probability. This method simplifies the modeling process by assuming independence between features, making it computationally efficient and effective for various classification tasks.

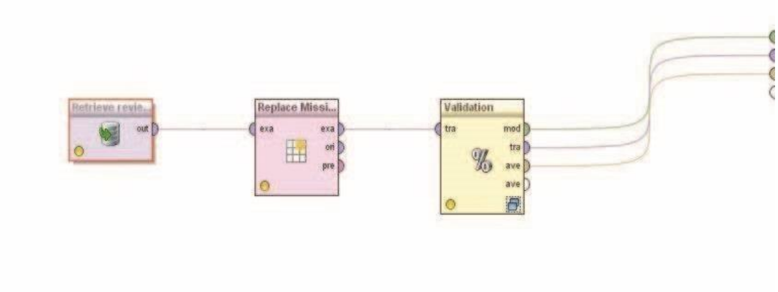


Fig: Illustrates the entire process work flow.

It shows three operators namely retrieve, replace missing values, Validation. The retrieve operator retrieves the dataset that is being uploaded in the tool. The replace the missing values operator replaces missing values if any. Replacement can be done by four methods namely minimum, maximum, average and zero. In order to estimate the statistical performance of a learning operator a cross-validation is performed by the validation operator.

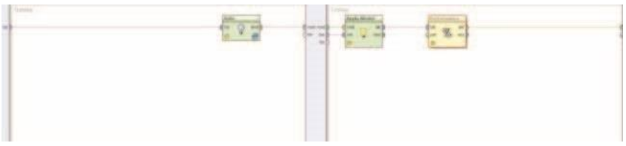


Fig : Illustrates the sub-process of cross validation operator.

The training process consists of the voting operator which is the technique that we propose for better results. On the testing sub process lies the apply model and performance operators which evaluate the correctness of the model.

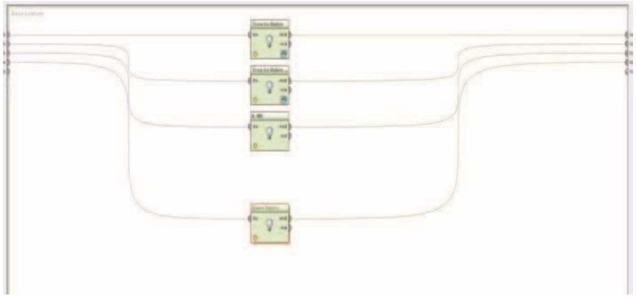


Fig : Illustrates the base learners which lie under the vote operator.

The system incorporates four distinct machine learning models: Naïve Bayes, K-Nearest Neighbor, CHAID, and Random Tree. Each of these machine learners is associated with specific operators, strategically positioned to perform their respective classification tasks. These operators execute the classification processes in alignment with the characteristics and requirements of each machine learning model. Additionally, the tree to rules operator is employed to derive rules directly from the CHAID and Random Tree models, facilitating a rule-based approach to better interpret and understand the decision-making process of these models.

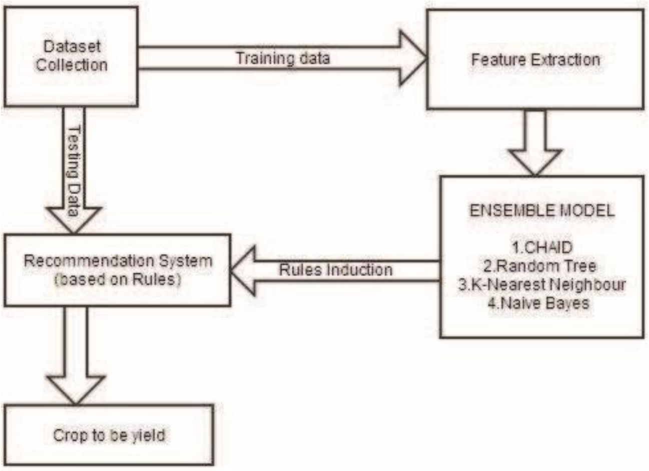


Fig : depicts the overall methodology of proposed system.

Rules induced from the Model:

IF pH is mild alkaline

AND depth is above 100

AND water holding capacity is LOW

AND drainage is moderately well

AND erosion is moderate

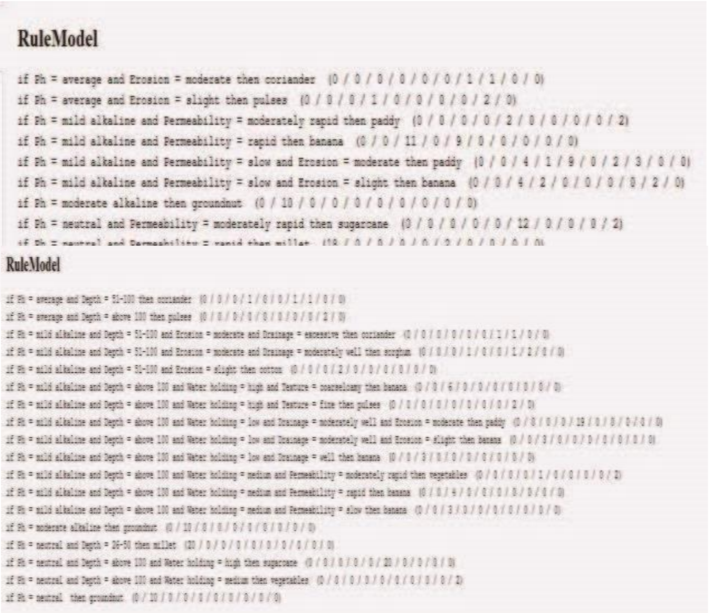
THEN recommend PADDY cultivation

This rule articulates the specific soil conditions necessary for cultivating the recommended crop, which, in this instance, is PADDY. The criteria outlined in the IF part of the rule represent the targeted soil attributes, guiding users in making informed decisions about suitable crops based on their specific soil characteristics.

RESULTS AND DISCUSSION

The model exhibits a commendable prediction accuracy, reaching 88%. The rules derived from both the Random Tree and CHAID models are presented in Figures 6 and 7, respectively. These rules adopt an if-then format, with the then part specifying the assigned class label. The precise classification results for each instance in the training set are clearly depicted.

The rules extracted from the aforementioned models, as illustrated in Figures 7 and 8, are instrumental in developing a Recommendation System. This system is realized through the creation of a Graphical User Interface (GUI), showcased in Figure 8, which is deployed as a web portal. The model, trained with the provided dataset, undergoes testing with user inputs via the portal. The implemented scripting responds to each test case, suggesting an appropriate crop based on the generated rules. In cases where the test input does not align with any predictions, a "no match" output is produced.



RuleModel

```
if Ph = average and Erosion = moderate then coriander  (0 / 0 / 0 / 0 / 0 / 0 / 1 / 1 / 0 / 0)
if Ph = average and Erosion = slight then pulses  (0 / 0 / 0 / 1 / 0 / 0 / 0 / 0 / 2 / 0)
if Ph = mild alkaline and Permeability = moderately rapid then paddy  (0 / 0 / 0 / 0 / 2 / 0 / 0 / 0 / 0 / 2)
if Ph = mild alkaline and Permeability = rapid then banana  (0 / 0 / 11 / 0 / 9 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Permeability = slow and Erosion = moderate then paddy  (0 / 0 / 4 / 1 / 9 / 0 / 2 / 3 / 0 / 0)
if Ph = mild alkaline and Permeability = slow and Erosion = slight then banana  (0 / 0 / 4 / 2 / 0 / 0 / 0 / 0 / 2 / 0)
if Ph = moderate alkaline then groundnut  (0 / 10 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = neutral and Permeability = moderately rapid then sugarcane  (0 / 0 / 0 / 0 / 0 / 12 / 0 / 0 / 0 / 2)
if Ph = neutral and Permeability = rapid then millet  (18 / 0 / 0 / 0 / 0 / 2 / 0 / 0 / 0 / 0)
```

RuleModel

```
if Ph = average and Depth = 51-100 then coriander  (0 / 0 / 0 / 1 / 0 / 0 / 1 / 1 / 0 / 0)
if Ph = average and Depth = above 100 then pulses  (0 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 2 / 0)
if Ph = mild alkaline and Depth = 51-100 and Erosion = moderate and Drainage = excessive then coriander  (0 / 0 / 0 / 0 / 0 / 0 / 1 / 1 / 0 / 0)
if Ph = mild alkaline and Depth = 51-100 and Erosion = moderate and Drainage = moderately well then sorghum  (0 / 0 / 0 / 1 / 0 / 0 / 1 / 2 / 0 / 0)
if Ph = mild alkaline and Depth = 51-100 and Erosion = slight then cotton  (0 / 0 / 0 / 2 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = high and Texture = coarseloamy then banana  (0 / 0 / 6 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = high and Texture = fine then pulses  (0 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 2 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = low and Drainage = moderately well and Erosion = moderate then paddy  (0 / 0 / 0 / 0 / 19 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = low and Drainage = moderately well and Erosion = slight then banana  (0 / 0 / 3 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = low and Drainage = well then banana  (0 / 0 / 3 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = medium and Permeability = moderately rapid then vegetables  (0 / 0 / 0 / 0 / 1 / 0 / 0 / 0 / 0 / 2)
if Ph = mild alkaline and Depth = above 100 and Water holding = medium and Permeability = rapid then banana  (0 / 0 / 4 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = mild alkaline and Depth = above 100 and Water holding = medium and Permeability = slow then banana  (0 / 0 / 3 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = moderate alkaline then groundnut  (0 / 10 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = neutral and Depth = 26-50 then millet  (20 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
if Ph = neutral and Depth = above 100 and Water holding = high then sugarcane  (0 / 0 / 0 / 0 / 0 / 20 / 0 / 0 / 0 / 0)
if Ph = neutral and Depth = above 100 and Water holding = medium then vegetables  (0 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 2)
if Ph = neutral  then groundnut  (0 / 10 / 0 / 0 / 0 / 0 / 0 / 0 / 0 / 0)
```

Fig 6 Rules induced from random tree model.

CONCLUSION

India, being a nation where agriculture holds a paramount position, the prosperity of the nation is intricately tied to the well-being of its farmers. Our efforts are dedicated to supporting farmers in making informed decisions, specifically in choosing the right seeds based on soil requirements. By doing so, we aim to contribute to increased productivity and profitability for farmers through precision agriculture techniques. This approach enables farmers to plant the most suitable crops, thereby enhancing individual yields and, collectively, elevating the overall agricultural productivity of the nation. Our future endeavors focus on refining our methodology by working with an improved dataset, incorporating a larger number of attributes, and introducing yield prediction capabilities. These enhancements aim to provide farmers with even more valuable insights and contribute to the sustainable development of agriculture in the country.

REFERENCES

[1].Aakunuri Manjula, Dr.G .Narsimha (2015), ‘XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction’ , Conference on Intelligent Systems and Control (ISCO)

[2].Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh (2015), ’Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique’, International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM).

[3].Aymen E Khedr, Mona Kadry, Ghada Walid (2015), ‘Proposed Framework for Implementing Data Mining Techniques to Enhance Decisions in Agriculture Sector Applied Case on Food Security Information Center Ministry of Agriculture, Egypt’, International Conference on Communications, management, and Information technology (ICCMIT').

[4].Roshani Ade, P.R.Deshmukh (2014), ‘Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice’, International Conference on Information and Communication Technologies (ICICT).

[5].Saso Karakatic, Marjan Hericko and Vili Podgorelec (2015), ‘Weighting and sampling data for individual classifiers and bagging with genetic algorithms’ International Joint Conference and Computational Intelligence(IJCCI).