Abstract—Machine learning has arisen with advanced data analytics. Many factors influence crop yield, such as soil, amount of water, climate, and genotype. Determining factors that significantly influence yield prediction and identify the most appropriate predictive methods are important in yield management. It is critical to consider and study the combination of different crop factors and their impact on the yield. The objectives of this paper are: (1) to use advanced data analytic techniques to precisely predict the soybean crop yields, (2) to identify the most influential features that impact soybean predictions, (3) to illustrate the ability of Fuzzy Rule-Based (FRB) sub-systems, which are self-organizing, self-learning, and data-driven, by using the recently developed Autonomous Learning Multiple-Model First-order (ALMMo-1) system, and (4) to compare the performance with other well-known methods. The ALMMo-1 system is a transparent model, which stakeholders can easily read and interpret. The model is a data-driven and composed of prototypes selected from the actual data. Many factors affect the yield, and data clouds can be formed in the feature/data space based on the data density. The data cloud is the key to the IF part of FRB sub-systems, while the THEN part (the consequences of the IF condition) illustrates the yield prediction in the form of a linear regression model, which consists of the yield features or factors. In addition, the model can determine the most influential features of the yield prediction online. The model shows an excellent prediction accuracy with a Root Mean Square Error (RMSE) of 0.0883, and Non-Dimensional Error Index (NDEI) of 0.0611, which is competitive with state-of-the-art methods.

Keywords—Autonomous Learning, Fuzzy Rule-Based, Multi-Modal, Linear regression, Soybean yield prediction

I. INTRODUCTION

Estimation of agricultural crop yields and production is of paramount importance for global food security. Crop yields result from the interaction of multiple factors such as soil and weather conditions, water availability, crop species, and management practices (sowing date, planting rate, etc.). However, the relative relevance of the different factors varies strongly from place to place and from season to season [21]. Understanding the effect of these factors on crop yields is crucial for guiding farmer's decisions to optimize farming practices and outputs in each region. Although high quality data on these factors is becoming increasingly available at the plot level [22], its analysis remains challenging [23]. Machine learning can be used as advanced data analytics to predict crop yields and determine the influence of the various factors on prediction accuracy in an intelligible and understandable manner, thereby providing support for farm decisions. Thus, machine learning can be used to promote successful agriculture and tackle the challenges in agricultural production. Extracting knowledge from data facilitates intelligent decision-making by humans. Over the last two decades, many machine learning algorithms were introduced that provide good prediction of linearly separable classes [1]. However, processing large and complex data streams requires non-linearly separable classes. Numerous non-linear methods are used to solve complex problems related to data streams [2, 3]. The Fuzzy Rule-Based (FRB) system is considered a universal approximator [4], and the neuro-fuzzy system is an instance of FRB, integrating an Artificial Neural Network (ANN) and an FRB system [5], by making use of ANN’s layered structures to represent a set of fuzzy rules. It could be interpreted as a multi-layer feed-forward neural network, or as a degree of likelihood in a probability distribution [2]. In 2018, the ALMMo-1 system, which is an extended development of a neuro-fuzzy system was proposed. The model uses an empirical approach for data partitioning with the robust design of an FRB system. It has a data-driven mechanism that learns autonomously from the underlying patterns of data streams [5]. ALMMo-1 can be a competitive proposed solution for the task in use. It is an online evolving mechanism that does not need any prior assumptions or parameters and can be easily adapted to the environment, due to the associated self-learning and self-organizing mechanisms. The objective of this paper is the following: to use ALMMo-1’s method to precisely predict soybean yields, from a database that combines environmental information, agricultural management and remote sensing information from commercial plots of Central Argentina. Furthermore, we determine the coefficient between factors from underlying patterns of data, thereby identifying the most important factor affecting soybean yield prediction, and compare the performance of the model with other well-known methods.

In simple terms, the ALMMo-1 model transforms the inputs, which are factors that affect the soybean yield, to the corresponding output, which is the prediction of soybean yield as shown in fig.1.
The model during the learning process using the antecedent of proxy of the crop condition). Vegetation Index (EVI) registered during the crop cycle as a vegetation index" (the maximum value of the Enhanced accounts for the weather conditions) and "the maximum amount" (total rainfall between 1-Nov and 31-Mar, which "plant density" (the number of plants per square meter), "sowing date" (expressed as the number of days from Jan.1st), in Central Argentina: three management factors which include identified five important factors that determine soybean yields growth conditions for each plot and season, respectively. We predicting factors, to account for the weather and vegetation Imaging Spectroradiometer (MODIS) [28] as important rainfall and remote sensing data from the Moderate Resolution AACREA [27]. In addition to the provided data, we included (2005/06 to 2017/18), by the private farmer’s association commercial plots in Central Argentina over 13 crop seasons soybean yield and management practices recorded for outputs. In this paper, we analyse a database containing determination, to optimize farming practices and production understanding of the multiple factors involved in yield this crop to the economy of the country demands a better different environmental conditions [26]. The central role of the country, with ~19 Mha harvested per year across very biodiesel) [25]. In addition, soybean is the main crop grown in and the largest exporter of soybean derivatives (cake, oil and supplier being the third largest soybean exporter in the world, good source of vegetable protein for the human diet and as a raw material in the production of biodiesel [24]. Argentina is an important soybean supplier being the third largest soybean exporter in the world, and the largest exporter of soybean derivatives (cake, oil and biodiesel) [25]. In addition, soybean is the main crop grown in the country, with ~19 Mha harvested per year across very different environmental conditions [26]. The central role of this crop to the economy of the country demands a better understanding of the multiple factors involved in yield determination, to optimize farming practices and production outputs. In this paper, we analyse a database containing soybean yield and management practices recorded for commercial plots in Central Argentina over 13 crop seasons (2005/06 to 2017/18), by the private farmer’s association AACREA [27]. In addition to the provided data, we included rainfall and remote sensing data from the Moderate Resolution Imaging Spectroradiometer (MODIS) [28] as important predicting factors, to account for the weather and vegetation growth conditions for each plot and season, respectively. We identified five important factors that determine soybean yields in Central Argentina: three management factors which include “sowing date” (expressed as the number of days from Jan. 1st), “plant density” (the number of plants per square meter), “maturity group” (the length of the crop lifecycle); “rainfall amount” (total rainfall between 1-Nov and 31-Mar, which accounts for the weather conditions) and “the maximum vegetation index” (the maximum value of the Enhanced Vegetation Index (EVI) registered during the crop cycle as a proxy of the crop condition).

II. AGRICULTURAL/ ECONOMIC RELEVANCE

Soybean is one of the most valuable crops in the world, not only as an oil seed crop and feed for livestock, but also as a good source of vegetable protein for the human diet and as a biofuel feedstock [24]. Argentina is an important soybean supplier being the third largest soybean exporter in the world, and the largest exporter of soybean derivatives (cake, oil and biodiesel) [25]. In addition, soybean is the main crop grown in the country, with ~19 Mha harvested per year across very different environmental conditions [26]. The central role of this crop to the economy of the country demands a better understanding of the multiple factors involved in yield determination, to optimize farming practices and production outputs. In this paper, we analyse a database containing soybean yield and management practices recorded for commercial plots in Central Argentina over 13 crop seasons (2005/06 to 2017/18), by the private farmer’s association AACREA [27]. In addition to the provided data, we included rainfall and remote sensing data from the Moderate Resolution Imaging Spectroradiometer (MODIS) [28] as important predicting factors, to account for the weather and vegetation growth conditions for each plot and season, respectively. We identified five important factors that determine soybean yields in Central Argentina: three management factors which include “sowing date” (expressed as the number of days from Jan. 1st), “plant density” (the number of plants per square meter), “maturity group” (the length of the crop lifecycle); “rainfall amount” (total rainfall between 1-Nov and 31-Mar, which accounts for the weather conditions) and “the maximum vegetation index” (the maximum value of the Enhanced Vegetation Index (EVI) registered during the crop cycle as a proxy of the crop condition).

III. THE PROPOSED COMPUTATIONAL APPROACH

The proposed data-driven approach uses an FRB framework with the recently introduced ALMMo model [5]. The model has many robust characteristics with which it makes decisions: learning from the stream data, self-adapting to environmental changes, self-monitoring from the data stream [2], [6], [7]. The FRB system is a multi-model system with the ability to learn autonomously and requiring neither human experts nor prior assumptions [6]. In this paper, the model predicts the yield of soybeans based on the densities of underlying data represented by the yield’s factors. In general, there are two types of ALMMo models: the zero-order ALMMo-0 and the first order ALMMo-1. The former, is used for classification, while the latter (considered in this paper) is used for regression. The structure of the model consists of data clouds, which represent the building blocks of the multi-model system [5]. Meta-parameters are generated from data clouds, which are formed completely from observed data without either user-specified or problem-specified parameters, and they are updated recursively. The model has sophisticated characteristics, which distinguish it from other methods because it is built based on Empirical Data Analysis (EDA), which employs local and global typicality, cumulative proximity, square centrality, and density [8]. Due to the recursive update of data, the model can run with a few samples of data, which promotes the computational efficiency and reduces memory consumption [5]. In the following subsections, more details about the approach will be presented.

A. Normalisation

Normalisation is used to adjust the values of each feature input into a specific range. In statistics, there are many types of normalization. In this paper, minimum and maximum feature scaling is used according to the following expression [1]:

$$x_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

(2)

The key notation used in this paper is described in table I.

After applying the above equation, the values of each feature lie in the range \([0,1]\).

B. The Architecture of ALMMo-1 model

The architecture of the ALMMo-1 model is essentially composed of multiple models, which are used in adaptive, control, prediction, and classification [6]. Multi-model systems follow the principle of “divide and rule” [2, 6, 9] by decomposing complex problems into a set of simpler ones and combining these afterwards. Examples of this model include, but are not limited to, FRB systems with a set of AnYa type local sub-systems initially proposed as an alternate to Mamdani and Takagi–Sugeno types. These FRB systems follow the actual data distribution and do not involve any aggregation or scalar membership functions because the data cloud density replaces the scalar membership function. These data clouds represent the data density distribution of observed data and derive the fuzzy membership of a particular IF… THEN of FRB. The construction of the IF part
consists of focal points that represent the data clouds. While, the consequent THEN part defines the local linear model [10],[5]. Each local linear model is obtained from an AnYa type fuzzy rule with prototype \( p_\mu \) and parameter vector \( a_\mu \) in the consequence part as follows [5]:

\[
\text{IF } (x \sim p_\mu) \text{ THEN } (y_\mu = a_\mu x^T) \quad (3) \\
y = \sum_{\mu=1}^{N} a_\mu x^T \quad (4)
\]

The key novelty of the ALMMo-1 model is the use of the EDA framework. The EDA depends entirely on the actual data; therefore, it is free from parameters, and assumptions [5],[11]. The main characteristics that have been used in the proposed model are unimodal discrete density \( D [8] \) and multimodal continuous typicality \( t^{MM} [12] \). \( \mathbb{R}^{M} \) is the set of the actual samples in the data stream and is denoted \( \{x_i\} = \{x_1, x_2, ..., x_N\} \), where the subscript \( S \) indicates the time instance. Unimodal discrete density \( D \) is generated from the ensemble properties and distributions of actual data samples and identifies the distance of a particular data sample in relation to all other data samples. The Euclidean distance is the most common type of distance used. The definition of the unimodal discrete density of the data sample \( x \) \( (x(x) \in \{x\}S \) formed from Cauchy function is [8]:

\[
D_S(x) = \frac{1}{1 + \frac{\|x - x_S\|^2}{\sigma^2_S}} \quad (5)
\]

The average scalar product \( X_S \) and the mean \( \mu_S \) updates recursively as follows [2]:

\[
\mu_S = \frac{s-1}{s} \mu_{S-1} + \frac{1}{s} x_S; \quad \mu_1 = x_1 \quad (6) \\
X_S = \frac{s-1}{s} X_{S-1} + \frac{1}{s} \|x_S\|^2; \quad X_1 = \|x_1\|^2 \quad (7)
\]

1) The identification of the antecedent (IF) part

When each data cloud arrives, known as \( x_{S+1} \), the model updates the global mean \( \mu_S \) and the average scalar product \( X_S \) using (6) and (7) to obtain \( \mu_{S+1} \) and \( X_{S+1} \), respectively. Then, the unimodal discrete density \( D \) and the focal points of the current data cloud \( \mu_{S+1} \) for the fuzzy rule \( z \) is updated using equation (5), denoted by \( D_{S+1}(x_{S+1}) \) and \( D_{S+1}(\mu_{S+1}) \). The first condition checks if a new rule will be generated as Cond.1:

\[
\text{IF } (D_{S+1}(x_{S+1}) > \max_{z=1,2,...,N} \left(D_{S+1}(\mu_{S,z})\right)) \text{ OR IF } (D_{S+1}(x_{S+1}) < \min_{z=1,2,...,N} \left(D_{S+1}(\mu_{S,z})\right)) \text{ THEN } (x_{S+1} \text{ is a new focal point}) \quad (8)
\]

When the first condition is met, a new focal point for the new data cloud is identified. Then, the second condition examines whether the new data cloud overlaps as Cond.2:

\[
\text{IF } (D_{S+1}(x_{S+1}) \geq \frac{1}{\sqrt{\pi}} \text{THEN} (the \ Nth \ focal \ point \ and \ the \ respective \ data \ cloud \ needs \ to \ be \ replaced \ by \ a \ new \ one) \quad (9)
\]

In the case of the first and second conditions are not fulfilled, the model creates a new rule with a new associated focal point \( x_{S+1} \), and the model initialises other parameters. However, when the first and second condition are fulfilled, the intersecting data cloud is removed and replaced by a new rule with a new associated focal point \( N_{S+1} \). When the first condition is not satisfied, \( x_{S+1} \) merges to the closest data cloud. The activation level of the \( z^\text{th} \) fuzzy rule at the \( S^\text{th} \) time instance is \( \lambda_{S,z} \), which is defined as the normalised unimodal density per rule as follows:

\[
\lambda_{S,z} = \frac{D_{S,z}(x_S)}{\sum_{z=1}^{N} D_{S,z}(x_S)} \quad (10)
\]

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>The feature input before normalisation, ( x^T \in \mathbb{R}^x ).</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>The feature input after normalisation, ( x_1 = [x_{1,1}, x_{1,2}, ..., x_{1,M}]^T \in \mathbb{R}^M ), ( i = 1, 2, ..., S ) and ( M ) is the features number.</td>
</tr>
<tr>
<td>( \min(x) )</td>
<td>The minimum value of all the data input in the dataset, ( i = [1, 2, ..., S] ).</td>
</tr>
<tr>
<td>( \max(x) )</td>
<td>The maximum value of all the data input in the dataset, ( i = [1, 2, ..., S] ).</td>
</tr>
<tr>
<td>( p_\mu )</td>
<td>The ( z^\text{th} ) data cloud that represents a fuzzy rule, and ( z = 1, 2, ..., N ); in this paper ( N=16 ).</td>
</tr>
<tr>
<td>( y_\mu )</td>
<td>The output of the ( z^\text{th} ) fuzzy rule.</td>
</tr>
<tr>
<td>( x^T )</td>
<td>The parameters of the linear model, ( x^T = [1, x^{2-1}] ) = ([1, x_{i,2}, ..., x_{i,M}]).</td>
</tr>
<tr>
<td>( a_\mu )</td>
<td>The coefficient of each (consequent) parameter of the ( z^\text{th} ) fuzzy rule.</td>
</tr>
<tr>
<td>( a_{\mu,z} )</td>
<td>The coefficient of each (consequent) parameter of the ( z^\text{th} ) fuzzy rule at instance ( S ) and ( a_{\mu,z} = [a_{\mu,1}, a_{\mu,2}, ..., a_{\mu,M}]^T ).</td>
</tr>
<tr>
<td>( \lambda_{S,z} )</td>
<td>The activation level of the ( z^\text{th} ) fuzzy rule.</td>
</tr>
<tr>
<td>( \lambda_{S,z+1} )</td>
<td>The activation level of the ( z^\text{th} ) fuzzy rule at instance time ( S + 1 ).</td>
</tr>
<tr>
<td>( D_S )</td>
<td>The unimodal discrete density of the data set ( {x}_S ).</td>
</tr>
<tr>
<td>( \mu_S )</td>
<td>The mean of the data samples in ( S ) dataset.</td>
</tr>
<tr>
<td>( \sigma_S^2 )</td>
<td>The deviation is a measure of difference between the actual value of a variable and other data samples on the data set ( {x}_S ). ( \sigma_S^2 = X_S -</td>
</tr>
<tr>
<td>( N_{S+1} )</td>
<td>( N ) is a number of rules at time ( S ), when a data sample arrives. ( N_{S+1} ) is a new rule and focal point ( x_{S+1} ).</td>
</tr>
<tr>
<td>( X_S )</td>
<td>The average scalar product.</td>
</tr>
<tr>
<td>( c_{S,z+1} )</td>
<td>The covariance matrix at time ( S + 1 ) of the fuzzy rule ( z ) in the consequence parameter.</td>
</tr>
<tr>
<td>( c_{S,N_{S+1}} )</td>
<td>The covariance matrix at ( S ) and the focal point ( N ) at time ( S+1 ) in the consequence parameter.</td>
</tr>
<tr>
<td>( x_{S+1} )</td>
<td>The input feature at time ( S + 1 ), ( x_1 ), is the input feature at time ( S ).</td>
</tr>
</tbody>
</table>

2) The identification of consequence (THEN) part

The system updates the consequence parameters. When a new rule is generated, the consequence parameters are established as follows:

\[
\begin{align*}
& a_{S,N_{S+1}} \leftarrow \frac{1}{N} \sum_{z=1}^{N_S} a_{S,z} \\
& C_{S,N_{S+1}} \leftarrow \Omega (M+1) \times (M+1)
\end{align*}
\]
The new rule inherits the consequence parameters $C_{S_{z+1}}$ and $a_{S_{z+1}}$ of the old rule, when both the first and second conditions are met. The omega parameter $\Omega$ is a constant parameter, and is set to 1000 in this paper. To optimize the consequent parameters of each rule, and the fuzzy weighted recursive least squares is used [13] as follows:

$$C_{S_{z+1}} = C_{S_{z}} - \frac{\lambda_{S_{z+1}} c_{S_{z}} x_{S_{z}+1} x_{S_{z}+1}^T c_{S_{z}}}{1+\lambda_{S_{z+1}} x_{S_{z}+1} x_{S_{z}+1}^T c_{S_{z}}} (13)$$

$$a_{S_{z+1}} = a_{S_{z}} + \lambda_{S_{z+1}} C_{S_{z+1}} x_{S_{z}+1} (y_{S_{z}+1} - x_{S_{z}+1}^T a_{S_{z}}) (14)$$

### IV. EXPERIMENTAL RESULTS AND EVALUATION

The dataset used in this experiment relates to agricultural data of soybean crop yields. The most influential factors affecting soybean yields are used as inputs to predict the yield and produce an autonomous learning multi-model regression model based on FRB. The importance of the method is that the model allows the stakeholders to navigate into the inner model representation and interpret the outcomes. The model has linguistic IF... THEN statements, which are interpretable, understandable, and explainable [29]. In addition, the stakeholders can identify and recognise the different factors that affect the yield. Therefore, they can control the yield based on the factors, “sowing data”, “density”, “maturity group”, “amount of rain”, and “the maximum of vegetation index”. For example, the yield can be increased based on the factors used in the FRB model that influence the crop. The main contributions of this paper are the use of the ALMMo-1 model in soybean yield prediction and achieving high prediction accuracy. The performance of a model in the learning and validation phases, which are competitive with state-of-the-art methods, are demonstrated in table II.

<table>
<thead>
<tr>
<th>Rule #</th>
<th>The expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IF $x_{5.1} \sim 169.82$ THEN $y = 0.069 + 4.65x_{5.1} + 0.713x_{5.2} + -0.0269x_{5.3} + 1.952x_{5.4} + 0.334x_{5.5}$</td>
</tr>
<tr>
<td>2</td>
<td>IF $x_{5.1} \sim 165.85$ THEN $y = 0.02 + 1.66x_{5.1} + 0.37x_{5.2} + 0.02x_{5.3} + 2.27x_{5.4} + 0.32x_{5.5}$</td>
</tr>
<tr>
<td>3</td>
<td>IF $x_{5.1} \sim 165.72$ THEN $y = 0.14 + 2.59x_{5.1} + 0.62x_{5.2} + 0.03x_{5.3} + 1.88x_{5.4} + 0.18x_{5.5}$</td>
</tr>
</tbody>
</table>

#### A. The formulation of IF... THEN FRB rule

The model that was used generating an identification of fuzzy rules during the learning process from the input features or factors, by creating data clouds formed from many data densities. A prototype of data density represents a component of one fuzzy rule, which consists of five parameters representing the soybeans factors (“sowing date”, “plant density”, “maturity group”, “rainfall amount”, and “the maximum vegetation index”) represented by $x_{5.1}, x_{5.2}, x_{5.3}, x_{5.4}, x_{5.5}$, respectively. In this paper, 16 rules were generated. Three examples of these rules are shown in table III.

### TABLE III. EXAMPLES OF FUZZY RULES FORMED FROM LEARNING PROCESS OF DATA CLOUDS

<table>
<thead>
<tr>
<th>Rule #</th>
<th>The expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IF $x_{5.1} \sim 169.82$ THEN $y = 0.069 + 4.65x_{5.1} + 0.713x_{5.2} + -0.0269x_{5.3} + 1.952x_{5.4} + 0.334x_{5.5}$</td>
</tr>
<tr>
<td>2</td>
<td>IF $x_{5.1} \sim 165.85$ THEN $y = 0.02 + 1.66x_{5.1} + 0.37x_{5.2} + 0.02x_{5.3} + 2.27x_{5.4} + 0.32x_{5.5}$</td>
</tr>
<tr>
<td>3</td>
<td>IF $x_{5.1} \sim 165.72$ THEN $y = 0.14 + 2.59x_{5.1} + 0.62x_{5.2} + 0.03x_{5.3} + 1.88x_{5.4} + 0.18x_{5.5}$</td>
</tr>
</tbody>
</table>

#### B. The prediction of soybean yield

This section presents the performance of the model prediction. Fig. 2 shows samples of data ordered from the minimum to the maximum yield. This demonstrates the convergence between the actual and predicted values of the soybean yield.

---

**TABLE II. The Performance of The Model in The Learning and Validation Phases**

<table>
<thead>
<tr>
<th>Phases</th>
<th>Correlation</th>
<th>RMSE</th>
<th>NDEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning</td>
<td>0.8408</td>
<td>0.0130</td>
<td>0.0621</td>
</tr>
<tr>
<td>Validation</td>
<td>0.9142</td>
<td>0.0883</td>
<td>0.0611</td>
</tr>
</tbody>
</table>

---

**Fig. 2. Percentage of each feature contribution in IF part of FRB.**

**Fig. 3. Comparing actual and predicted yields.**
The percentage of each feature contribution in the antecedent IF part is shown in fig. 3. The features from 1 to 5 indicate “sowing date”, “plant density”, “maturity group”, “rainfall amount”, and “the maximum vegetation index”, respectively.

C. Comparison between ALMMo-1 model and other algorithms

Comparison between the ALMMo-1 model and other popular algorithms is shown in table IV. The model achieves excellent results compared to other well-known algorithms.

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMMo-1</td>
<td>0.0883</td>
</tr>
<tr>
<td>Ensemble-Bagged Trees</td>
<td>0.0935</td>
</tr>
<tr>
<td>Tree-Course Tree</td>
<td>0.0994</td>
</tr>
<tr>
<td>SVM- Coarse Gaussian SVM</td>
<td>0.1064</td>
</tr>
<tr>
<td>Linear regression- Robust Linear</td>
<td>0.1183</td>
</tr>
<tr>
<td>Stepwise Linear Regression</td>
<td>0.1120</td>
</tr>
</tbody>
</table>

In addition, the non-dimensional error index (NDEI) is used to evaluate the performance of the proposed method. The NDEI expression is given by:

$$NDEI = \sqrt{\frac{\sum (\gamma_{\text{actual}} - \gamma_{\text{estimated}})^2}{\sigma_{\text{actual}}}}$$  \hspace{1cm} (15)

Where $\gamma_{\text{estimated}}$ is a value of the prediction and corresponds to the system output, $\gamma_{\text{actual}}$ is the actual truth-value, and $\sigma_{\text{actual}}$ is the standard deviation of the actual value. The performance of the proposed algorithm using NDEI is 0.0611.

V. ALTERNATIVE METHODS

Many methods exist to predict crop yield in precision agriculture [14, 15], [16]. The most commonly used methods for predicting crop yields are correlation and multiple regression models. These methods are used to identify essential features affecting the yield. Non-linear statistical methods have been introduced to investigate yield estimation [17], using simple linear correlations of yield with soil properties. However, there are variations between fields and years. Many studies used ANN to predict yields [18], [14]. For example, in a study on setting target corn yields, the root mean square error (RMSE) was 20% [18]. In a comparison of the efficiency of multiple linear regression models and ANN models to predict the soybean yields in Maryland, the ANN model performed better, with an $r^2$ and RMSE of 0.81 and 214 compared with 0.46 and 312 for the linear regression model [14]. Other methods have also been used. For example, a study of potato yield prediction via satellite image and precision agriculture techniques, employed Vegetation Indices (VIs), which include NDVI and SAV. The method gained an accuracy of between 7.9 and 13.5% for Landsat-8 images and between 3.8 and 10.2% for Sentinel-2 images [19]. A few studies have been proposed with regards to knowledge-based systems using fuzzy logic techniques in agricultural domain [15], [16], and some authors proposed a framework of a fuzzy rule-based cognitive map [20], [17].

VI. CONCLUSION

In this paper, the ALMMo-1 system is implemented to predict soybean crop yields from factors that affect the yield. The model achieves high accuracy. Distinctive characteristics of the model are the interpretability and readability that allows stakeholders to navigate through the internal model representation, therefore, the factors that impact soybean yields can be adjusted. Also, the model uses nonparametric data clouds, which are assumption-free, constraint-free, totally data-driven, and recursively updated. The model evolves and updates with each data sample entry, improving memory and computation efficiency. In future work, the method could be extended to process the crop images, map yields with satellite data, and evaluate the resulting yield predictions.

REFERENCES


