Bioresource Technology Reports 22 (2023) 101421

Contents lists available at ScienceDirect

Bioresource Technology Reports

journal homepage: www.sciencedirect.com/journal/bioresource-technology-reports

Toward sustainable crop residue management: A deep ensemble learning approach

Syeda Nyma Ferdous [a,*], Xin Li [a,*], Kamalakanta Sahoo [b], Richard Bergman [b]

[a] Lane Department of Computer Science and Electrical Engineering, West Virginia University, WV 26506, Morgantown 26507, WV, USA
[b] Forest Products Laboratory, Forest Service, United States Department of Agriculture, Madison 53726, WI, USA

ARTICLE INFO

Keywords:
Biomass sustainability indicators
Deep ensemble model
Soil conditioning index
Organic matter factor
Soil erosion factor

ABSTRACT

For the existence of biorefineries, a consistent supply of sustainable crop residues is critical. Crop residue removal may adversely affect the overall sustainability, i.e., soil productivity and economics of crop residue removal. Therefore, it is necessary to develop a tool or method to estimate the sustainable removal of crop residues which is critical to various stakeholders including farmers, biorefinery, and policymakers. We proposed a robust deep ensemble machine learning(ML) model to estimate soil sustainability indicators [i.e., soil-erosion-factor(SEF), soil-conditioning-index(SCI), and organic-matter-factor(OMF) as well the sustainable crop residues removal rates (RRR) using topography, soil properties, climate, and crop management practices. The results showed that proposed ML models are having very high predictability [SEF,$R^2 = 0.999$; SCI,$R^2 = 0.996$; OMF,$R^2 = 0.996$, and RRR,$R^2 > 0.98$) with low error for sustainability indicators and RRR. The proposed ML model can be used as a tool for estimating sustainable RRR and serve as a guide to assess sustainability indicators in real-time.

1. Introduction

The United Nations Climate Change Conference in Glasgow (COP26) has raised alarm about the dire consequences of climate change due to the increase in greenhouse gas (GHG) emissions in the Earth's atmosphere. Countries have agreed to move away from fossil fuels and switch to low-carbon energy sources. Globally, bioenergy accounts for over 10 % of the total primary energy demand. Bioenergy is one of the low-carbon energy sources that can significantly reduce GHG emissions and mitigate climate change. Plant biomass, including crop residues, forest residues, and energy crops, is one of the most prominent bioenergy resources (Energy Agency, 2021; Energy Information Administration, n.d.; Langholtz et al., 2016a). Crop residues are one of the most abundant and readily available renewable biomass resources for the production of bioenergy. Complete removal of crop residues from agricultural land can negatively impact soil quality and land productivity (Blanco-Canqui and Lal, 2009; Lal, 2005; Turmel et al., 2015). Therefore, an optimal and predetermined amount of crop residues can be removed from agricultural land without adversely affecting the soil, which can be described as sustainable crop residues management or production.

Sustainable crop residue production is highly dependent on the accurate estimation of different environmental indicators (Bridgwater, 2006)considering the conditions of a farm and crop management activities. The most important and widely used sustainability indicators to define the sustainable removal of crop residues from agricultural land are the soil erosion factor (SEF), the soil conditioning index (SCI), and the organic matter factor (OMF) related soil of cropland. These factors are used to express soil quality, which is an important consideration for quantifying the availability of sustainable crop residues in a study area. Indicators that define sustainable removal of crop residues, such as topography, soil properties, crop management practices, and crop yield, vary between farms and in a study region from which a biorefinery draws the biomass. Therefore, a reliable and accurate assessment of sustainably available crop residues at the field level is required to reduce the risks of insufficient biomass availability, to design an optimal supply chain network, and to suggest long-term policy decisions. Several studies have been conducted on the evaluation of sustainable crop residues to assist in the development of biorefinery plans using biophysical and empirical models and geochemical models. However, most of them use aggregated topography and other field-level variables at the county level, increasing uncertainty and reducing the reliability of the biomass assessment results, especially specific to farms in a studied region. Machine learning models can be used to accurately quantify different

* Corresponding authors.
E-mail addresses: sf0070@mix.wvu.edu (S.N. Ferdous), Xin.Li@mail.wvu.edu (X. Li).

https://doi.org/10.1016/j.biteb.2023.101421
Received 22 December 2022; Received in revised form 29 March 2023; Accepted 31 March 2023
Available online 8 April 2023

sustainability indicators over current approaches so that informed decisions can be made about biomass production in spatially explicit areas (Elmaz et al., 2020).

Machine learning approaches can potentially predict biomass estimations from field-level datasets for a large region that are becoming more frequently available. Gleason and Lm (Gleason and Im, 2012) conducted a study on the estimation of forest biomass using machine learning approaches. To predict the estimation of biomass for gasification, machine learning methods have been explored (Ali et al., 2015; Elmaz et al., 2020). Han et al. applied machine learning-based techniques to model above-ground biomass (Han et al., 2019). Wang et al. (2016) estimated wheat biomass using a machine-learning model. Sahoo et al. estimated cotton stalks in the southeast of the United States using an artificial neural network (ANN) model considering the local climate, topography, soil, and crop management practices (Sahoo et al., 2016). Sahoo et al. developed geographic information system (GIS) based predictive models to estimate sustainable crop residues and identified optimal locations of biorefinery (Sahoo et al., 2018). Machine learning models were explored by (Dodo et al., 2022)to predict the energy content of biomass. A detailed study by Morais et al. on grassland estimation using machine learning methods suggested that the use of machine learning in conjunction with remote sensing data can improve the performance of biomass estimation (Morais et al., 2021). In (Fathololoumi et al., 2022), a decision-based fusion strategy has been developed for the modeling of crop residues. Applications of machine learning for biomass-derived materials (BDM) have been studied by (Wang and Yao, 2023).

Assessing biomass for a large study area and using grid-level (30 × 30 meter) data in the GIS platform is time-consuming and memory-consuming. Machine learning models are a great alternative to designing a prediction system, as the time complexity and memory requirements are significantly reduced for these models while maintaining excellent accuracy. The optimal location and long-term reliability of the biorefinery are also affected by the sustainable availability of biomass. Therefore, our proposed approach can help investors and policymakers in the decision-making process for the future deployment of biorefineries in the United States and globally.

Ensemble Learning is a machine learning technique that is used to combine several base models into one for robust prediction performance (Mendes-Moreira et al., 2012). Generally, the performance of ensemble learning is better than individual models as individual models might not capture the features necessary for robust prediction. Although classical machine learning models such as linear regression (LR) (Su et al., 2012), multi-layer perceptron (MLP) (Ramchoun et al., 2016), tree-based algorithms such as decision tree (DT) (Myles et al., 2004), random forest (RF) (Breiman, 2001), eXtreme Gradient Boosting (XGBoost) (Chen and He, n.d.) based predictive models can be used to model biomass sustainability prediction problem, integration of deep and machine learning models can learn a generalized feature space making deep ensemble a natural choice for this task.

This proposed USDA-funded research study is part of a larger effort known as MASBio, the Mid-Atlantic Sustainable Biomass for Value-Added Products Consortium (https://masbio.wvu.edu/). The mid-Atlantic region of the US has abundant natural resources that can support a growing, yet sustainable bio-economy. This project aims to use integrated modeling and predictive analytics to estimate the landscape-level environmental impacts and costs of residue removal rates, perform scale-up analyses, and build a decision support system. Experimental data from various locations in the study area will be used to develop and validate biogeochemical models for plant biomass including hybrid willow, switchgrass, and forest residues. Machine learning tools will be used to generate high-spatial life-cycle inventories and the costs of producing multiple bioproducts. The main goal of this phase is to support the creation of an integrated economic and environmental decision support tool that utilizes big data landscape analytics for bioproducts produced on a plant-level scale.

In particular, the results of this predictive analysis on crop residues could support future high-resolution and landscape-level indicator estimates for sustainable residue removal rates for the three plant biomasses listed above in the mid-Atlantic region. To achieve this goal, we develop a deep ensemble model that combines a deep learning model with a robust machine learning model. The training of deep models can be complex. To stabilize the training and improve the generalization performance of the ensemble model, we applied orthogonal regularization in the hidden layer of deep models. Model performance was verified using standard goodness-of-fit metrics, such as cross-correlation (R2), root mean squared error (RMSE), and mean absolute error (MAE).

The objectives of this study were to (i) explore the potential of machine learning models in the context of predicting values of different biomass sustainability indicators considering the existing soil properties, topography, crop management practices, and climate for a single farm or farms in a large region such as state or country, (ii) propose a novel deep ensemble learning-based approach that can be utilized for the assessment of sustainable crop residues in a region, and (iii) evaluate the proposed model so that it can be reliably applied to platforms related to real-life biomass sustainability prediction.

2. Methods

2.1. Case study

In this study, the applicability of the machine learning model was investigated for the main sustainability indicators of crop residue removal from agricultural lands. An efficient deep ensemble model has been developed to estimate sustainability indicators relevant to crop residues harvesting(e.g. SEF, SCI, and OMF) with the highest correlation coefficient (R2) and the lowest standard error value. The analysis was performed on a field-level data set collected from the state of Ohio, which is heterogeneous and contains numerical and categorical data. The dataset was constructed by simulating synthetic data for sustainability indicators (SE, SCI, and OMF) using RUSLE2 software (RUSLE2, n.d.) from 23 counties in the state of Ohio and representing 165 types of soil. The dataset contains around 4500 data points spread across the state. Annual crop yield data at the county level was collected from the National Agricultural Statistics Service (NASS) (USDA-NASS, n.d.) during 2010–2014. Furthermore, soil-related information (e.g., soil type, texture, slope length, etc.) was collected from the Geographic Gridded Soil Survey (gSSURGO) database (Gridded Soil Survey Geographic Database (gSSURGO), n.d.). The simulation was performed for major soil types considering two types of crop rotations (crop rotation for corn-soybean, crop rotation for corn-soybean-wheat), no-tillage option, three levels of crop yield (low, average, and high) and five removal rates (no removal, 22 %, 35 %, 52 %, and 83 %). The input variables to predict sustainability indicators included biomass yield (corn, soybean, wheat), soil characteristics (soil erodibility, soil texture, soil organic matter content), rainfall erosivity, slope (soil surface inclined relative to horizontal), slope length, K-factor (microbial biomass potassium), crop rotations (corn-soybean, corn-soybean-wheat), and residue removal rate. The availability of sustainable crop residue depends on the studied input features.

2.2. Data pre-processing

The dataset needs to be a good representation of the corresponding feature set to train a robust model. We standardized the feature set by removing the mean and scaling to unit variance. Also, multi-collinearity is a well-known problem that hinders the stable training of a regression model (Farrar and Glauber, 1967; Graham, 2003). To handle the multicollinearity problem, we removed highly correlated independent variables that may have caused overfitting of the model.

From Fig. 1, we can observe that the dummy variables were created for two categories of crop rotations (Crop Rotations_CS for corn-soybean

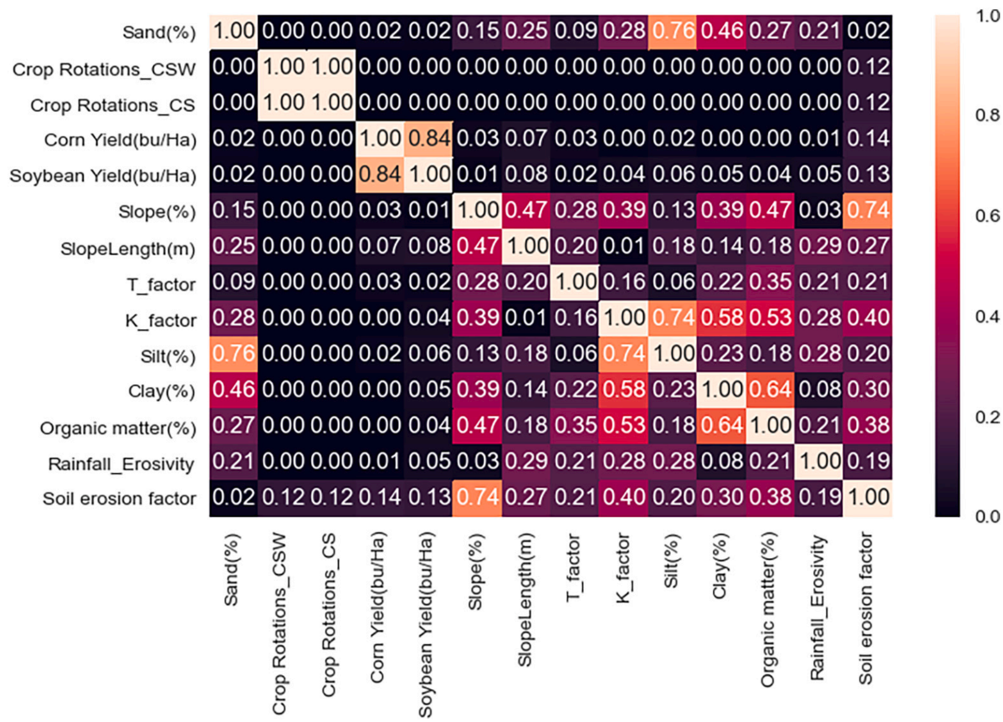

Fig. 1. Cross-correlation matrix showing different correlations between features.

crop rotation and Crop Rotations_CSW for corn-soybean-wheat crop rotation) respectively. So, we dropped Crop Rotations CS to consider only Crop Rotations CSW while generating the feature set for the model. The correlation coefficient (r_{xy}) was computed using Eq. (1):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}} \tag{1}$$

x_i, y_i respectively denote sample means and $\overline{x_i}$, $\overline{y_i}$ are data points and n is the sample set size.

2.3. Model design

This study presented a novel deep ensemble learning-based approach for the estimation of primary sustainability indicators of biomass production. Our proposed framework used stacking (Džeroski and Ženko, 2004) which is an ensemble machine-learning algorithm that uses a meta-learning method to learn how to best combine the predictions of machine-learning and deep-learning models. The benefit of stacking is that it can harness the capabilities of several well-performing models on either classification or regression tasks, which achieves predictions with better performance than any single model in the ensemble. The framework of our model is illustrated in Fig. 2.

2.3.1. Classical machine learning models

In this study, first, we analyzed the performance of classical machine learning models as a baseline to model the regression problem.

- **Linear regression (LR):** Linear regression (LR) is a supervised machine learning algorithm that assumes a linear relationship between a dependent variable and one or more independent variables. The model tries to predict the unknown parameters estimated from the data linearly (Seber and Lee, 2012; Su et al., 2012).
- **Multi-layer perceptron (MLP):** Multi-layer perceptron (MLP), one of the widely studied models for real-world data, is used to model nonlinear relationships underlying practical data These models are made up of nodes interconnected to each other through three different layers, namely the input layer, the hidden layer, and the output layer (Gardner and Dorling, 1998; Ramchoun et al., 2016). The input first passes through the input layer. Then, the hidden layer performs a series of transformations to compute the hidden representation of features. Finally, the output layer performs the specified task. MLP computes the weighted sum of the inputs and includes a bias for each training sample. The weighted total is passed through an activation function which is responsible for producing the final output.
- **k-nearest neighbors (KNN):** k nearest neighbors (kNN) can be used for classification and regression (Peterson, 2009). In both cases, the input consists of the k closest training examples in the data set. In a classification setting, the output is a class membership. An object is classified by a majority vote of its k nearest neighbors. The value of k can be predefined or can be assessed by cross-validation. In the regression setting, the output is the property value for the object obtained by the average of the values of k nearest neighbors. It uses different distance functions such as Euclidean, Manhattan, and Minkowski for distance measures between data points.
- **Support vector machine (SVM):** The Support Vector Machine (SVM) is a regression (supervised machine learning) algorithm that handles regression as a convex optimization problem (Basak et al., 2007; Noble, 2006). SVM tries to find the hyperplane by estimating the best-fit line which captures the maximum number of data points in a higher-dimensional feature space. The model tries to find the optimal hyperplane within a threshold value or error margin for improved performance. In SVM, the threshold value is the distance between the hyperplane and the boundary line. To map the data points in higher dimensions, it uses the kernel trick. Kernel trick uses different kernel functions (Gaussian kernel, radial basis function (RBF) kernel, etc.) for modeling optimal separation of nonlinear data points.
- **Decision tree (DT):** Decision trees (DTs), as the name implies, are tree-structured nonparametric supervised learning methods (Anthony J. Myles et al., 2004). These methods work by learning decision rules to solve both classification and regression problems. There

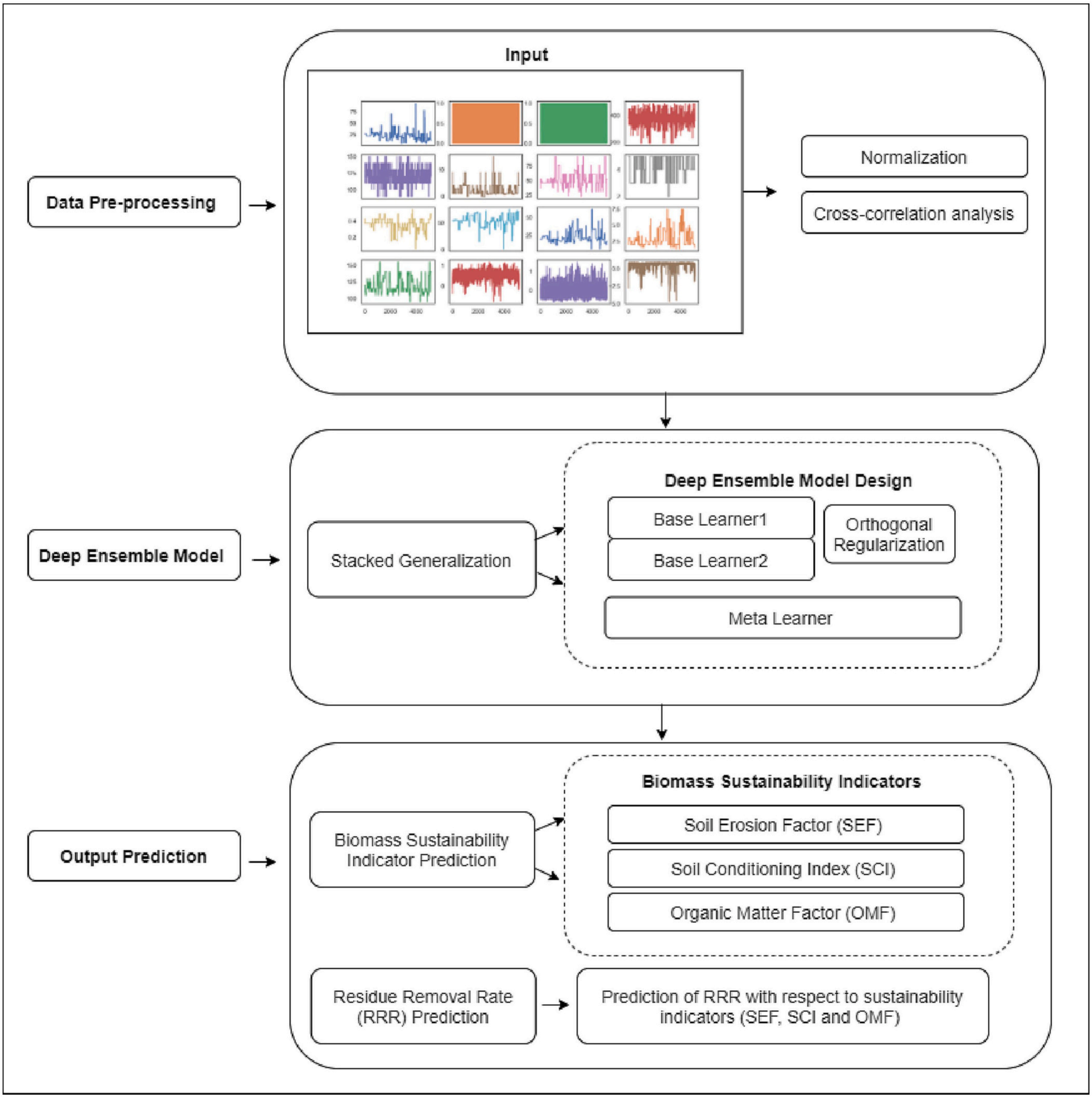

Fig. 2. Methodology framework of our proposed deep ensemble model.

are two types of nodes in a decision tree: internal nodes and leaf nodes where the internal nodes constitute the features corresponding to a dataset and the leaf nodes represent the class labels. The decision tree is split into branches that represent the classification rules used to make decisions for classifying objects and assign different class labels to them. To select the best features for splitting, some popular metrics (e.g., information gain, entropy, Gini impurity) have been devised.

- **Random forest (RF):** Random Forest (RF) is an ensemble-based supervised learning model in which a collection of uncorrelated decision tree models is trained in parallel using the bagging principle (Biau and Scornet, 2015). RF takes decisions considering all the combinations of the predicted results of individual decision tree outcomes, which helps to achieve good generalization performance in terms of solving regression problems. In RF, the training samples are selected randomly while performing splitting which prevents the overfitting problem. This algorithm performs well on large datasets and datasets with missing data points.
- **AdaBoost (ADB):** Adaptive Boosting, Adaboost; a widely used supervised boosting algorithm performs prediction using an iterative approach learning from the mistakes of the individual models sequentially (Schapire, n.d.). This algorithm combines predictions of multiple weak learners making a strong final model prediction.
- **XGBoost (XGB):** eXtreme Gradient Boosting (XGBoost) is another powerful machine-learning library that works based on a gradient boosting scheme (Chen and He, n.d.). In XGBoost, an ensemble of decision tree models is optimized using a gradient descent algorithm.

XG-Boost is preferred for its fast computational speed and high performance for predictive modeling.

Our model consisted of two levels: Level-0 and Level-1 respectively. Level-0 consists of the base learners, the models which are trained on k-fold cross-validation data and level-1 contains the meta-learner which is trained on the out-of-fold predictions. We have implemented two level-0 deep base learners demonstrated in Fig. 3. The first model used 1d convolution (conv1D) which was designed for sequential data processing (Visca et al., 2021). The second model used classical MLP (Ramchoun et al., 2016). The 1D convolution model had 8 layers having 5 1D CNN layers followed by a MaxPooling Layer. Then we used Flatten layer to get a 1D vector and passed it to the output layer for regression prediction. Our developed MLP network had 5 layers where the first one was an input layer followed by 4 hidden layers and an output layer. Each model was trained for 100 epochs with a learning rate of 0.0001. The orthogonal regularization coefficient λ was set to 0.01 initially and then gradually reduced to 0.001, 0.0001, and 0.00001 after every 20 epochs. We used XGBoost as a meta-learner as it best combines the deep models outperforming classical independent machine learning regressors for our dataset. The features extracted by the base models ensured diversity and generalization while training the meta-model and performing the final prediction.

2.4. Orthogonal regularization

Orthogonality is a desirable property for training deep networks as it helps to preserve energy activation within CNN layers by preventing issues related to network training such as vanishing gradient (Hochreiter, 1998), co-variate shift (Bickel et al., 2009) problem, and so on. Inspired by the successful application of orthogonal regularization (Bansal et al., 2018) on computer vision tasks such as image classification (Wang et al., 2020), image reconstruction (Brock et al., 2016) etc. we applied a modified strategy of orthogonal regularization to our proposed deep ensemble model as follows.

The orthogonality constraint ensured that the weights were orthogonal by pushing the weights to the nearest orthogonal manifold, minimizing the distance between the gram matrix and the identity matrix. Moreover, it improved feature quality and ensured stable convergence of deep models. We formulated the regularization using Eq. (2).

$$\lambda \left\| W^T W - I \right\|_F^2 \tag{2}$$

where λ is the regularization coefficient, W is the weight of the filter and I is the identity matrix. Note that the row-column orthogonality equivalence had been rigorously established by the lemma in Wang et al. (2020), which assured the unique regularization independent of the actual shape of W.

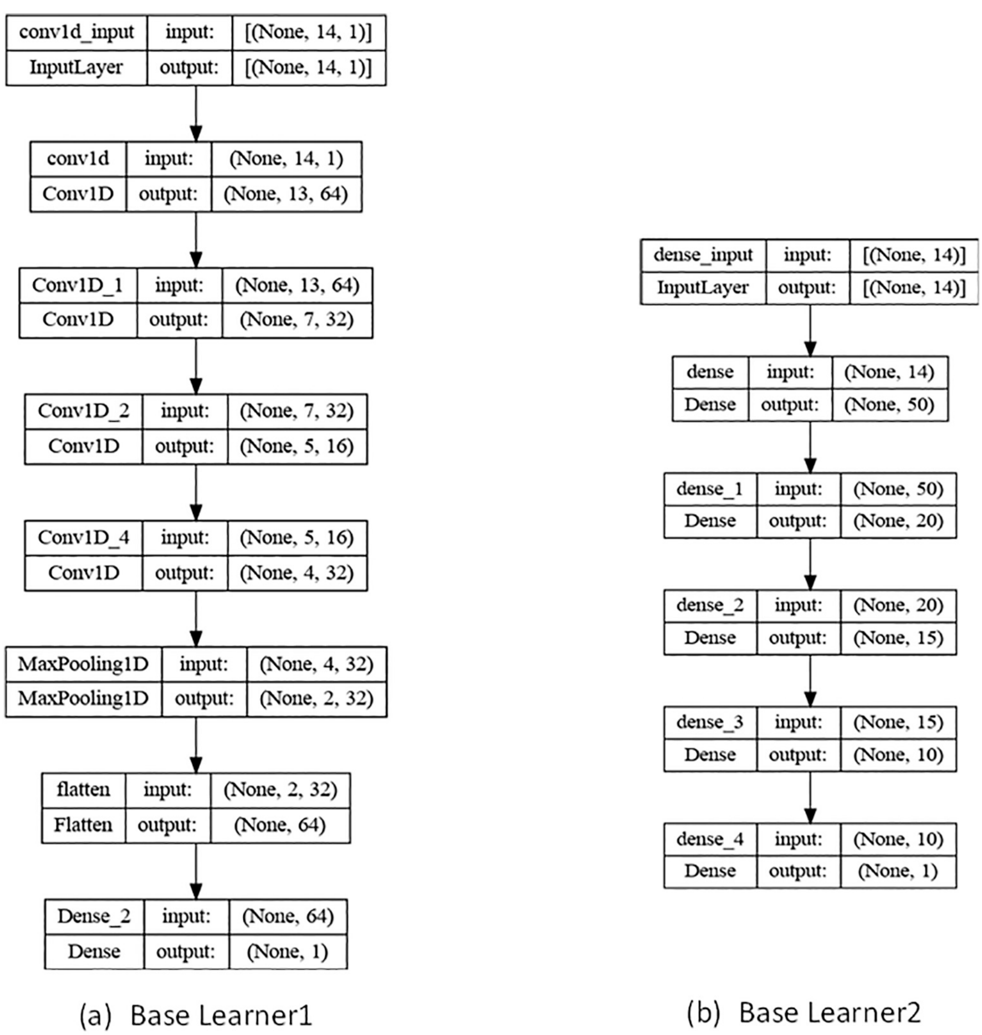

Fig. 3. Detailed model architecture of base learners.

2.5. Performance measure

The model performance was estimated based on R2, RMSE, and MAE metrics (Emmert-Streib and Dehmer, 2019). R^2 is the goodness-of-measure for how a model fits the data. R^2 was computed using Eq. (3). RMSE and MAE were computed using Eq. (4) and Eq. (5), respectively. In RMSE, errors are squared, giving relatively large weights to higher errors compared to MAE. Higher R^2 along with reduced MAE and RMSE ensures a good performance model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (S_a - S_p)^2}{\sum_{i=1}^{n} (S_a - \overline{S_a})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (S_p - S_a)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \|S_p - S_a\| \tag{5}$$

where n is the number of samples, S_a, S_p, S_a are the ground truth, predicted values, and actual mean values respectively. These metrics have been widely used in the existing literature of statistics and previous study on biomass assessment such as Sahoo et al. (2018).

3. Results and discussion

3.1. Estimation of sustainability indicators

The performance of the deep ensemble machine learning model to predict the soil erosion factor, soil conditioning index, and organic matter factor is reported in Table 1. Fig. 4 and Fig. 5 show the performance of the model for the prediction of the biomass sustainability indicator in terms of R2 metric and residual error. Fitted data points close to the line in Fig. 4 indicate that the model can accurately predict the sustainability indicators with low errors. Also, from Fig. 5, we can observe that the histogram of residual error plots followed almost normal distribution for the prediction of three biomass sustainability indicators which further support stable prediction of sustainability indicators by the developed models. The results in Fig. 4 and Fig. 5 suggest that the proposed method can be applied to the successful estimation of sustainable biomass production.

The trained models can also be used to assess the features that are important for biomass sustainability prediction, which is shown in Fig. 6.

Among all input parameters, the effect of the slope had the highest influence on the soil erosion factor. Analyzing the dataset, we found that an increase in slope decreases the soil erosion factor, implying a negative correlation between the two variables. Therefore, the slope is one of the most influential features for predicting soil erosion factors.

Crop residue removal adversely affects the overall sustainability which was defined by the SEF, SCI, and OMF. Removing crop residues enhances soil erosion but soil erosion kept under a threshold may not be detrimental to soil productivity (Muth and Bryden, 2013a; Muth and Bryden, 2013b; Langholtz et al., 2016a). Each unit of cropland (this study considered 30 × 30 m) may consider a calculated amount of crop residues removal rate that keeps the values of sustainability indicators below the prescribed values, i.e., the estimated values of SEF should be lower than the threshold value (usually 10 Mg/ha), the estimated SCI index should be positive, and the estimated OMF should be positive (Muth and Bryden, 2013b; Sahoo et al., 2018). It had been discussed that the farmers can remove crop residues from the cropland to keep the estimated values of these sustainability indicators below the threshold to main the productivity of the cropland (Langholtz et al., 2016b; Muth and Bryden, 2013b). The most influential input factors that influence the SEF are slope, crop residue removal rate, soil erodibility factor (k-factor), and rainfall erosivity. Avand et al. (2023) applied various ML models to estimate soil erosion from a watershed and conclude with similar findings as this study. The input parameters that had a significant impact on the SCI index were topography or slope, residue removal rate, soil properties (e.g. clay content), k-factor, and crop yield. Similarly, OMF is influenced by various input factors, especially soil properties, and residue removal rates. These findings are inconsistent with previous studies, especially the billion-ton report (Langholtz et al., 2016a). (Sahoo et al., 2018) demonstrated that the relationships between the input parameters and the sustainability indicators are mostly non-linear and therefore, the developed ensemble model could able to capture the nonlinear relationship with a very high predictability (>0.99).

Table 1
Performance of our model in predicting biomass sustainability indicators.

Model	R^2	MAE	RMSE
SEF	0.9988	0.0121	0.0032
SCI	0.9958	0.0092	0.0314
OMF	0.9965	0.0130	0.0578

3.2. Estimation of Residue Removal Rate (RRR)

This study identified several input indicators that influence the sustainability indicators such as SEF, SCI, and OMF. But among all the only input parameters that a user, especially a farmer has control over is the crop residue removal rates and this value is very useful to know before the harvesting of crop residues and production of sustainable crop residues for a biorefinery. Therefore, we developed an ensemble model to estimate the RRR for studied cropland with known values of three sustainability indicator factors, SEF, SCI, and OMF. Table 2 and Fig. 7 showed the performance of three ensemble models for the prediction of RRR. The predictability of these models is very high and with very low MAE and RSME.

Fig. 8 shows the effect of different features in RRR prediction concerning biomass sustainability indicators. The figure demonstrated slope, clay and organic matter factor as significant inputs for RRR prediction in terms of SEF, SCI and OMF. Sahoo et al. (Sahoo et al., 2018) performed a similar experiment in which they developed independent machine learning models using ANN to predict the residue removal rate for different biomass sustainability indicators in the same dataset. Unlike (Sahoo et al., 2018), we adopt an ensemble modeling approach where we leverage the benefits of prediction using multiple models. We compared our RRR prediction results with the best-performing ANN model reported by (Sahoo et al., 2018). From the results of Table 3, we can come to the conclusion that our developed model outperformed the proposed model in (Sahoo et al., 2018) for RRR prediction in terms of the three sustainability indicators of SEF, SCI, and OMF respectively. High R^2 and low error values showed that the proposed machine learning model is robust and reliable. Our models can mimic the results of empirical biophysical models such as RUSEL2 (RUSLE2, n.d.) which is widely used to estimate the availability of crop residues on agricultural land. It had been mentioned that biophysical empirical models such as RUSEL2 require high processing times and computing power if the estimation is performed for a large study area and with a high spatial resolution (i.e., 50 × 50 m) (Muth and Bryden, 2013a). We did not quantify the processing time required to estimate the RRR for a large study area. But based on the analysis of the requirement of processing time and resources for estimating the sustainable availability of crop residues at high spatial resolution (30 m × 30 m) and for a large study area by Sahoo et al. (2018), we can say that, Our proposed model may require a fraction of computing resources and time without compromising the accuracy of crop residue assessment. The adoption of precision agriculture technologies is becoming familiar to farmers due to increased productivity with lesser inputs such as nitrogen and thus improving the overall sustainability of agriculture (Denora et al., 2022;

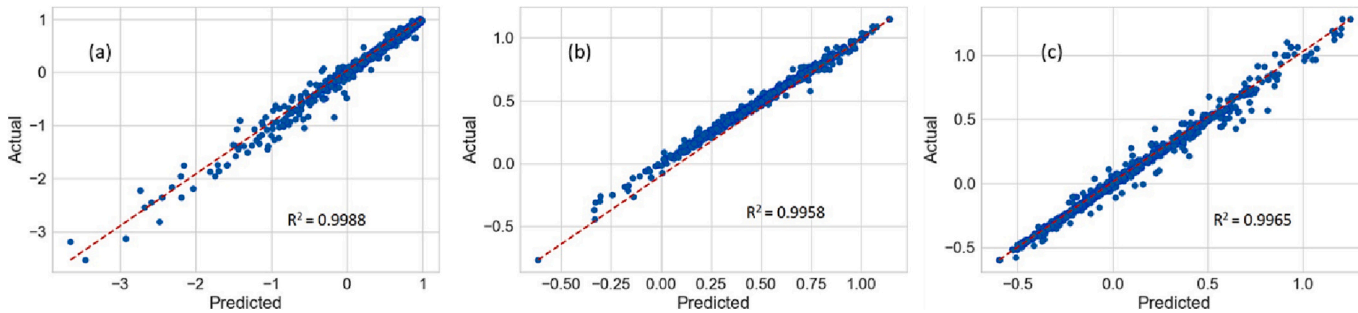

Fig. 4. Model performance using R^2 metric to predict biomass sustainability indicators.

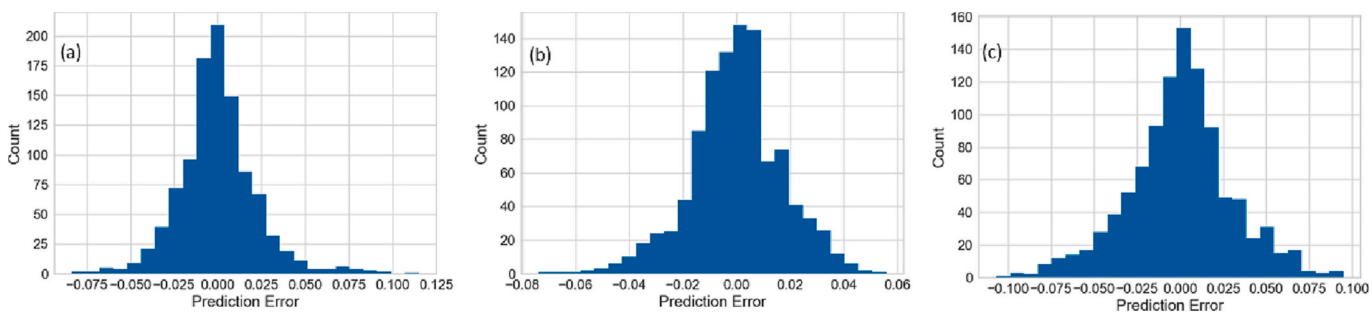

Fig. 5. Model performance in terms of residual error for biomass sustainability indicators (a) SEF. (b) SCI and (c) OMF.

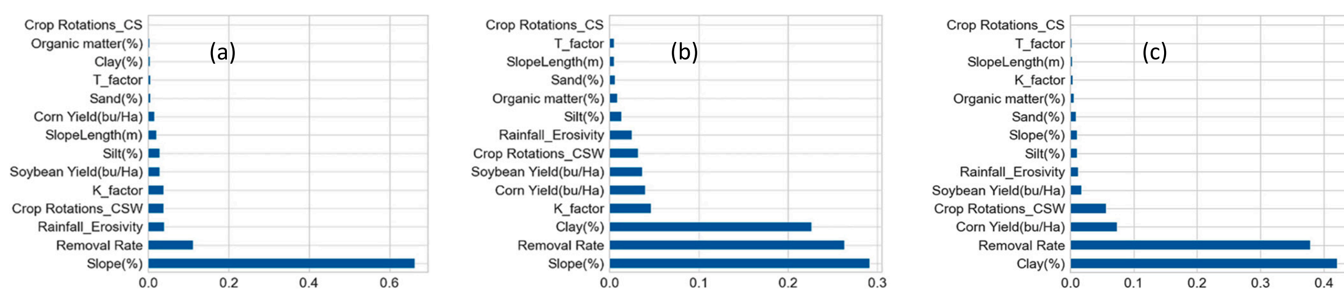

Fig. 6. Assessment of feature importance using the developed model for biomass sustainability indicator prediction (a) SEF (b) SCI and (c) OMF. Blue bars indicate the feature importance level. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Performance of our model to predict residue removal rate (RRR) for different sustainability indicators.

Model	R^2	MAE	RMSE
SEF	0.9846	0.0227	0.0032
SCI	0.9905	0.0180	0.0662
OMF	0.9911	0.0173	0.0443

Jin et al., 2019). Precision technology such as Variable rate techniques (VRT) requires prior field information at much higher spatial resolution and sometimes it may reduce to less than a meter to a few centimeters (Peter et al., 2020). Similar high resolution is expected in this type of study and the user can include inputs with much higher resolution (i.e., 1 × 1 m) information such as topography, soil properties, and crop yield for a large study area to get more accurate estimate of sustainable crop residues. The proposed model can act as an enabler in the implementation of precision agriculture for crops and quantify the removal of residue from agricultural lands specific to the individual plant in the field based on the information available at this scale.

3.3. Model convergence using orthogonal regularization

We observed that orthogonal regularization helped in training convergence of the developed models by minimizing the gaps between training and validation loss and stabilizes to a specific point. The loss curves shown in Fig. 9 correspond to the stable training of the model for the prediction of biomass sustainability.

From Table 4, we can see that adding orthogonal regularization to the baseline deep ensemble model decreased the RMSE. This performance improvement indicated that the addition of orthogonal regularization encourages the weights to lie close to the orthogonal manifold, which helps in effective feature generation for crop residue management.

3.4. Deep ensemble over independent machine learning models

We investigated the performance of classical machine learning models to predict biomass sustainability indicators and RRR using R^2, MAE, and RMSE metrics. Performance of LR (Su et al., 2012), MLP (Ramchoun et al., 2016), KNN (Peterson, 2009), SVM (Noble, 2006), DT (Myles et al., 2004), RF (Biau and Scornet, 2015), ADB (Schapire, n.d.) and XGB (Chen and He, n.d.) was reported for both tasks. Higher R^2 close to unity with lower RMSE and MAE metrics demonstrated the

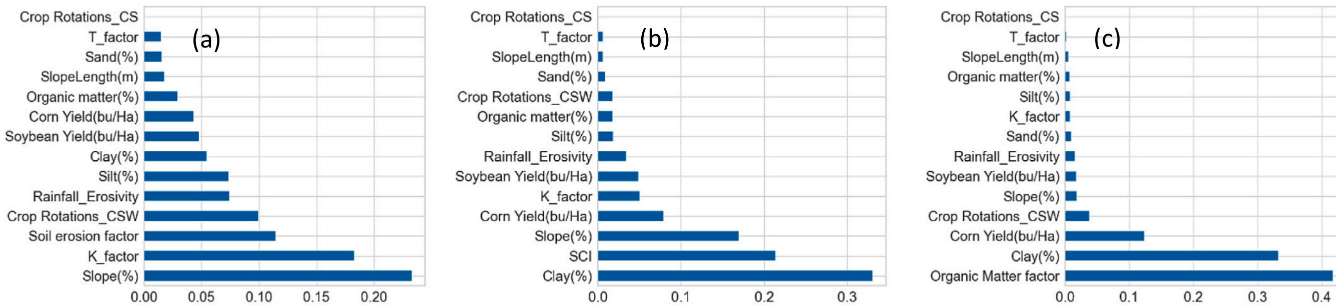

Fig. 7. Model performance using R^2 metric for predicting RRR with respect to (a) SEF (b) SCI and (c) OMF.

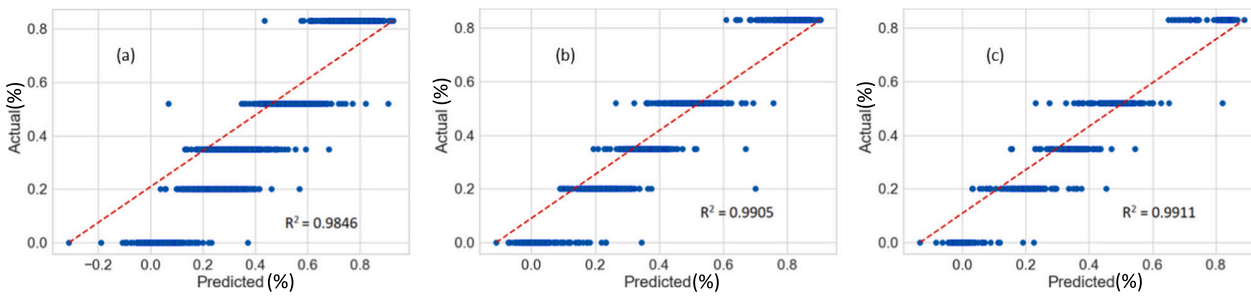

Fig. 8. Assessment of feature importance using the developed model for RRR prediction with respect to sustainability indicators (a) SEF (b) SCI and (c) OMF. Blue bars indicate the feature importance level. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of model performance using R^2 metric to predict the residue removal rate (RRR) for biomass sustainability indicators.

Model	SEF	SCI	OMF
(Sahoo et al., 2018)	0.88	0.95	0.74
This study	0.98	0.99	0.99

superiority of the developed models. Fig. 10 and Fig. 11 present individual machine-learning model performances to predict different biomass sustainability indicators and RRR. In both cases, XGBoost was the best-performing machine learning model. As shown in Fig. 12, our developed model performed stable prediction with a better generalization performance. Among classical independent machine learning models, XGBoost was the best-performing model for 80:20 and 70:30 dataset splitting, whereas MLP regression performed the best for 60:40 dataset splitting. Our proposed deep ensemble model performed the best with high R^2 values in varying split ratios. From Table 1 and Table 2, we observed that our proposed deep ensemble model outperformed the best-performing machine learning model in all scenarios. The high predictability of our developed model ensures that the nonlinearity of the response variables was better captured for the prediction of biomass sustainability.

4. Future works

ML is expected to accelerate the research on biomass and bioresource technologies in the following aspects:

1) Predictive modeling: ML algorithms can be used to build predictive models that help biomass researchers to identify potential sources of bioresources including both agricultural residues (e.g., crop stalks, straw, and husks) and energy crops (e.g., switchgrass, miscanthus, and willow). These models can be used to predict the yield and quality of crops, identify optimal harvesting times, and even forecast weather

Table 4
Error comparison using the RMSE metric to predict residue removal rate (RRR) with respect to biomass sustainability indicators.

Model	SEF	SCI	OMF
Baseline Deep Ensemble	0.0034	0.0663	0.0621
Baseline Deep Ensemble + Orthogonal Regularization	0.0032	0.0662	0.0443

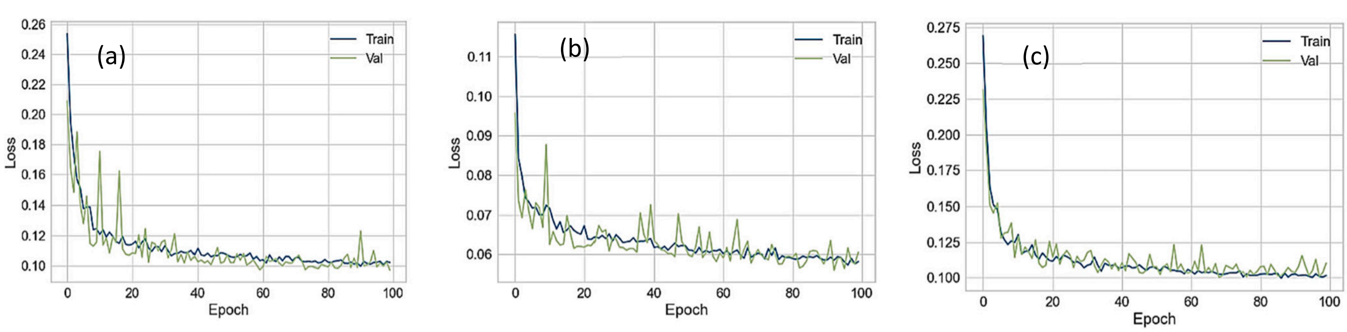

Fig. 9. Loss curves for training and validation data for the three sustainability prediction factors (a) SEF (b) SCI and (c) OMF.

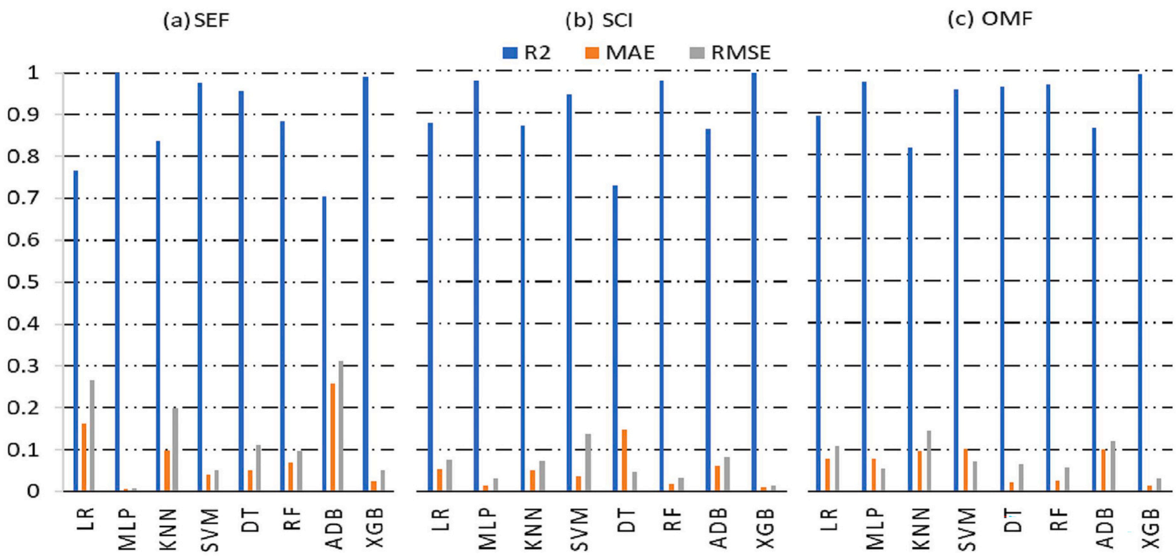

Fig. 10. Classical machine learning model performance for predicting biomass sustainability indicators (a) SEF (b) SCI and (c) OMF.

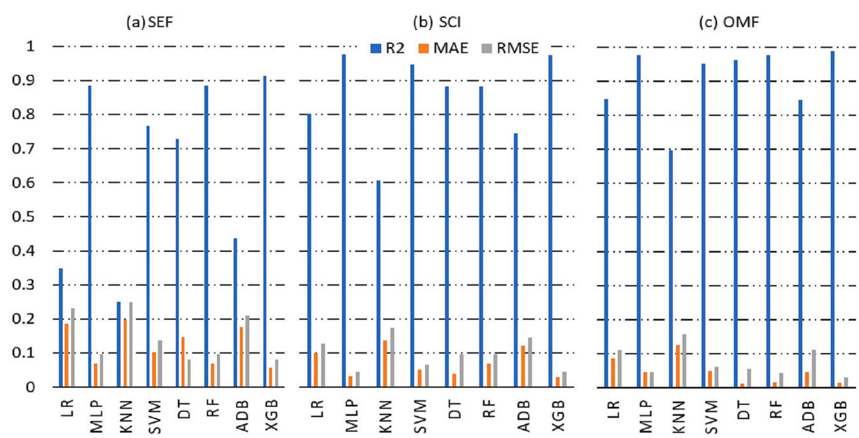

Fig. 11. Classical machine learning model performance for predicting RRR with respect to (a) SEF (b) SCI and (c) OMF.

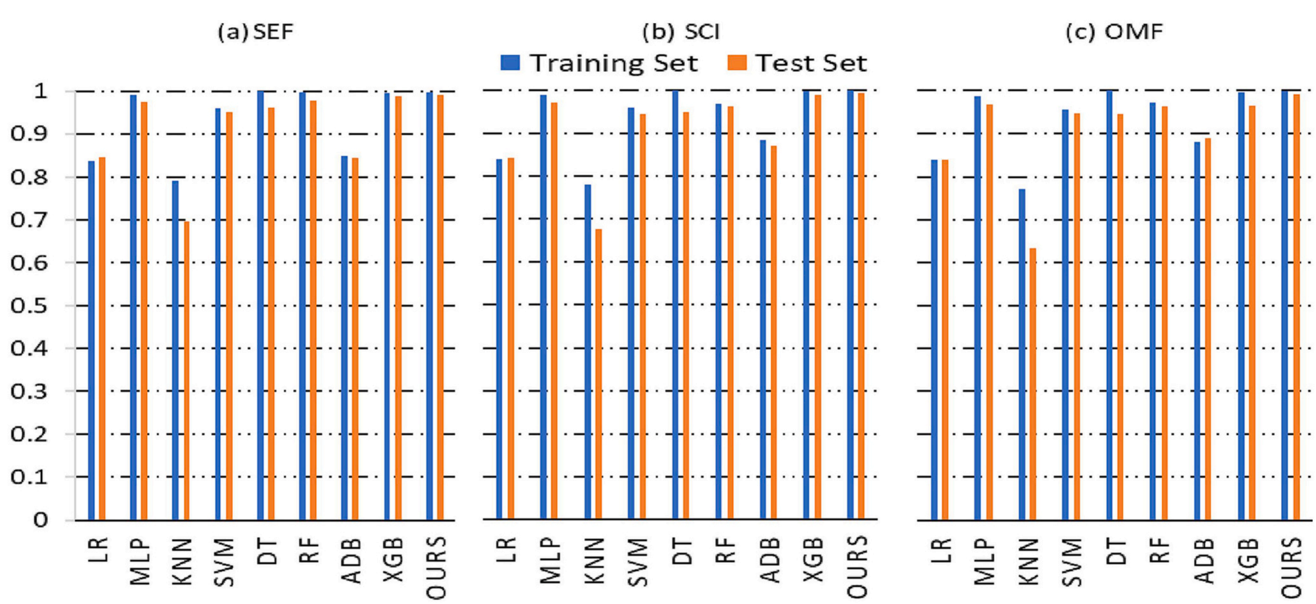

Fig. 12. Impact of splitting factor for predicting residue removal rate for organic matter factor (a) 80:20 (b) 70:30 and (c) 60:40.

patterns that affect crop growth. Our work is valuable to help stakeholders and policymakers in the decision-making process for the future deployment of biorefineries in the United States and globally.

2) Process optimization: ML algorithms can be used to optimize the processes involved in biomass and bioresource technology, such as fermentation and extraction. By analyzing data from these processes, ML models can identify the optimal conditions for maximum yield and quality. Currently, we are working with another task group of the MASBio project on optimizing the process of the concentrated and dilute acid hydrolysis of lignocellulosic biomass and related kinetic models.

3) Data analysis: ML algorithms can be used to analyze large datasets related to biomass and bioresources, such as genetic data, crop yields, and environmental data. This analysis can help researchers to identify patterns and relationships that would be difficult to discern using traditional statistical methods. In our future work, we plan to combine the deep ensemble model of this work with the DayCent model to accelerate the simulation of the cycling of carbon, nitrogen, and water in ecosystems over daily to decadal time scales.

5. Conclusion

This study demonstrates applications of machine learning models for the prediction of agricultural land sustainability indicators and the sustainable removal of crop residues from agricultural land using various inputs including topography, soil properties, climate, and crop management practices. The study findings revealed crop residue removal rates (RRR), slope and soil properties as significant inputs for the prediction of biomass sustainability indicators. A deep ensemble learning-based model applying stacked generalization was developed for this purpose, which outperformed classical machine learning models along with the previously developed model for this task. The developed model can be used to identify relevant features for sustainable crop residue management, as well as to assess sustainable crop residues removal rates accurately and reliably for large study areas using very high spatial resolution input data, i.e., 30mx30m or less such as 1×1 m. Our proposed model had standard goodness-of-fit, ensuring reliable performance for the given tasks particularly the prediction of three biomass sustainability indicators (SEF ($R^2 = 0.9988$), SCI ($R^2 = 0.9958$), and OMF ($R^2 = 0.9965$)) and sustainable RRR ((SEF ($R^2 = 0.9846$), SCI ($R^2 = 0.9905$), and OMF ($R^2 = 0.9911$)) respectively. The developed models can now be integrated with the GIS platform in real-time, so the model could be used as a live decision support tool for precision agricultural industry and quantifying sustainable RRR for the consistent supply crop residues to future biorefineries. Our next step is to apply the proposed deep ensemble model in real-world scenarios for validating its predictive power. The ML-based framework is also applicable to the acceleration of other biomass-related biogeochemical simulations and predictions (e.g., based on the DayCent model (Parton et al., 1998).

CRediT authorship contribution statement

Syeda Nyma Ferdous works on experiments and paper writing.

Xin Li works on paper writing and research supervision.

Kamalakanta Sahoo is responsible for experimental design and paper proofreading.

Richard Bergman is responsible for research collaboration and paper proofreading.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported by Agriculture and Food Research Initiative Competitive Grant no. 2020-68012-31881 from the USDA National Institute of Food and Agriculture. The findings and conclusions in this report are those of the author(s) and should not be construed to represent any official USDA or U.S. Government determination or policy. This research was supported in part by the USDA, Forest Service. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government.

References

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. https://doi.org/10.3390/rs71215841.

Avand, M., Mohammadi, M., Mirchooli, F., Kavian, A., Tiefenbacher, J.P., 2023. A new approach for smart soil erosion modeling: integration of empirical and machine-learning models. Environ. Model. Assess. 28, 145–160. https://doi.org/10.1007/s10666-022-09858-x.

Bansal, N., Chen, X., Wang, Z., 2018. Can we gain more from orthogonality regularizations in training deep networks? Adv. Neural Inf. Proces. Syst. 31.

Basak, D., Pal, S., Patranabis, D.C., 2007. Support Vector Regression,Neural Information Processing-Letters and Reviews. https://doi.org/10.1007/978-1-4302-5990-9_4.

Biau, G., Scornet, E., 2015. A Random Forest Guided Tour. https://doi.org/10.1007/s11749-016-0481-7.

Bickel, S., Brückner, M., Brückner, B., Scheffer, T., 2009. Discriminative learning under covariate shift. J. Mach. Learn. Res. 10, 2137–2155. https://www.jmlr.org/papers/volume10/bickel09a/bickel09a.pdf.

Blanco-Canqui, H., Lal, R., 2009. Crop residue removal impacts on soil productivity and environmental quality. Crit. Rev. Plant Sci. 28 (3), 139–163. https://doi.org/10.1080/07352680902776507.

Breiman, L., 2001. Random Forests. https://doi.org/10.1023/A:1010933404324.

Bridgwater, T., 2006. Biomass for energy. J. Sci. Food Agric. 86, 1755–1768. https://doi.org/10.1002/jsfa.2605.

Brock, A., Lim, T., Ritchie, J.M., Weston, N., 2016. Neural Photo Editing With Introspective Adversarial Networks. https://doi.org/10.48550/arXiv.1609.07093.

Chen, T., He, T. n.d. xgboost: eXtreme Gradient Boosting. https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf.

Denora, M., Amato, M., Brunetti, G., De Mastro, F., Perniola, M., 2022. Geophysical field zoning for nitrogen fertilization in durum wheat (Triticum durum Desf.). PLoS One 17 (4), e0267219. https://doi.org/10.1371/journal.pone.0267219.

Dodo, U.A., Ashigwuike, E.C., Abba, S.I., 2022. Machine learning models for biomass energy content prediction: a correlation-based optimal feature selection approach. Bioresour.Technol.Rep. 19, 101167 https://doi.org/10.1016/J.BITEB.2022.101167.

Džeroski, S., Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one? Mach. Learn. 54, 255–273.

Elmaz, F., Yücel, Ö., Mutlu, A.Y., 2020. Predictive modeling of biomass gasification with machine learning-based regression methods. Energy 191, 116541. https://doi.org/10.1016/J.ENERGY.2019.116541.

Emmert-Streib, F., Dehmer, M., 2019. Evaluation of regression models: model assessment, model selection and generalization error. Mach. Learn. Knowl. Extr. https://doi.org/10.3390/make1010032.

I. Energy Agency, 2021. World Energy Outlook 2021 [Online]. Available: www.iea.org/weo.

U. Energy Information Administration, "Annual Energy Outlook 2022 (AEO2022)." [Online]. Available: www.eia.gov.

Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in Regression Analysis: The Problem Revisited. https://doi.org/10.2307/1937887.

Fathololoumi, S., Karimi Firozjaei, M., Biswas, A., 2022. Innovative fusion-based strategy for crop residue modeling. Land (Basel) 11. https://doi.org/10.3390/land11101638.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

Gleason, C.J., Im, J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sens. Environ. 125, 80–91. https://doi.org/10.1016/J.RSE.2012.07.006.

Graham, M.H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84, 2809–2815. https://doi.org/10.1890/02-3114.

[database] Gridded Soil Survey Geographic Database (gSSURGO), https://data.nal.usda.gov/dataset/gridded-soil-survey-geographic-database-gssurgo., n.d.

Han, L., Yang, G., Dai, H., Xu, B., Yang, H., Feng, H., Li, Z., Yang, X., 2019. Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. Plant Methods 15. https://doi.org/10.1186/s13007-019-0394-z.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int.J.Uncertain/Fuzziness Knowl.-Based Syst. 6, 107–116. https://doi.org/10.1142/S0218488598000094.

Jin, Z., Archontoulis, S.V., Lobell, D.B., 2019. How much will precision nitrogen management pay off? An evaluation based on simulating thousands of corn fields over the US Corn-Belt. Field Crop Res. 240, 12–22. https://doi.org/10.1016/J.FCR.2019.04.013.

Lal, R., 2005. World crop residues production and implications of its use as a biofuel. Environ. Int. 31 (4), 575–584. https://doi.org/10.1016/j.envint.2004.09.005.

Langholtz, M.H., Stokes, B.J., Eaton, L.M., 2016a. 2016 Billion-Ton Report: Advancing Domestic Resources for a Thriving Bioeconomy. Jul. https://doi.org/10.2172/1271651.

Langholtz, M.H., Stokes, B.J., Eaton, L.M., 2016b. 2016 Billion-ton report: Advancing domestic resources for a thriving bioeconomy, Volume 1: Economic availability of feedstock. U.S. Department of Energy. ORNL/TM-2016/160. Oak Ridge National Laboratory, Oak Ridge, TN, p. 411.

Mendes-Moreira, J., Soares, C., Jorge, A.M., de Sousa, J.F., 2012. Ensemble approaches for regression: a survey. ACM Comput. Surv. https://doi.org/10.1145/2379776.2379786.

Morais, T.G., Teixeira, R.F.M., Figueiredo, M., Domingos, T., 2021. The use of machine learning methods to estimate above ground biomass of grasslands: a review. Ecol. Indic. 130, 108081 https://doi.org/10.1016/J.ECOLIND.2021.108081.

Muth, D.J., Bryden, K.M., 2013a. An integrated model for assessment of sustainable agricultural residue removal limits for bioenergy systems. Environ. Model Softw. 39, 50–69. https://doi.org/10.1016/j.envsoft.2012.04.006.

Muth, D.J., Bryden, K.M., 2013b. An integrated model for assessment of sustainable agricultural residue removal limits for bioenergy systems. Environ. Model Softw. 39, 50–69. https://doi.org/10.1016/j.envsoft.2012.04.006.

Myles, Anthony J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. https://doi.org/10.1002/cem.873.

Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. https://doi.org/10.1038/nbt1206-1565.

Parton, W.J., Hartman, M., Ojima, D., Schimel, D., 1998. DAYCENT and its land surface submodel: description and testing. Glob. Planet. Chang. 19 (1–4), 35–48.

Peter, B.G., Messina, J.P., Carroll, J.W., Zhi, J., Chimonyo, V., Lin, S., Snapp, S.S., 2020. Multi-spatial resolution satellite and sUAS imagery for precision agriculture on smallholder farms in Malawi. Photogramm. Eng. Remote. Sens. 86, 107–119. https://doi.org/10.14358/PERS.86.2.107.

Peterson, L.E., 2009. In: K-nearest Neighbor, 4. Scholarpedia, p. 1883. https://doi.org/10.4249/scholarpedia.1883.

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., Ettaouil, M., 2016a. Multilayer perceptron: architecture optimization and training. Int.J.Interact.Multimedia Artif. Intell. 4, 26. https://doi.org/10.9781/ijimai.2016.415.

Revised Universal Soil Loss Equation, Version 2 (RUSLE2), https://data.nal.usda.gov/dataset/revised-universal-soil-loss-equation-version-2-rusle2 [WWW Document], n.d.

Sahoo, K., Hawkins, G.L., Yao, X.A., Samples, K., Mani, S., 2016. GIS-based biomass assessment and supply logistics system for a sustainable biorefinery: a case study with cotton stalks in the Southeastern US. Appl. Energy 182, 260–273. https://doi.org/10.1016/j.apenergy.2016.08.114.

Sahoo, K., Mani, S., Das, L., Bettinger, P., 2018. GIS-based assessment of sustainable crop residues for optimal siting of biogas plants. Biomass Bioenergy 110, 63–74. https://doi.org/10.1016/j.biombioe.2018.01.006.

Schapire,n.d.Schapire, R.E., n.d. Explaining AdaBoost. doi:10.1007/978-3-642-41136-6_5.

Seber, G.A., Lee, A.J., 2012. Linear Regression Analysis. John Wiley & Sons. Springer-Verlag Berlin Heidelberg GmbH. https://doi.org/10.1002/9780471722199.

Su, X., Yan, X., Tsai, C.L., 2012b. Linear regression. Wiley Interdiscip. Rev. Comput. Stat. 4, 275–294. https://doi.org/10.1002/wics.1198.

Turmel, M.S., Speratti, A., Baudron, F., Verhulst, N., Govaerts, B., 2015. Crop residue management and soil health: a systems analysis. Agric. Syst. 134, 6–16. https://doi.org/10.1016/j.agsy.2014.05.009.

USDA-NASS, https://quickstats.nass.usda.gov/ [WWW Document], n.d.

Visca, M., Bouton, A., Powell, R., Gao, Y., Fallah, S., 2021. Conv1D Energy-Aware Path Planner for Mobile Robots in Unstructured Environments. https://doi.org/10.1109/ICRA48506.2021.9560771.

Wang, J., Chen, Y., Chakraborty, R., Yu, S.X., 2020. Orthogonal convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11505–11515.

Wang, H.S.H., Yao, Y., 2023. Machine learning for sustainable development and applications of biomass and biomass-derived carbonaceous materials in water and agricultural systems: a review. Resour. Conserv. Recycl. 190, 106847 https://doi.org/10.1016/J.RESCONREC.2022.106847.

Wang, L., Zhou, X., Zhu, X., Dong, Z., Guo, W., 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. 4, 212–219. https://doi.org/10.1016/j.cj.2016.01.008.