agriculture

MDPI

Article

Wild Blueberry Harvesting Losses Predicted with Selective Machine Learning Algorithms

Humna Khan [1], Travis J. Esau [1,*], Aitazaz A. Farooque [2] and Farhat Abbas [3]

[1] Department of Engineering, Faculty of Agriculture, Dalhousie University, Truro, NS B2N 5E3, Canada
[2] Faculty of Sustainable Design Engineering, University of Prince Edward Island, Charlottetown, PE C1A 4P3, Canada
[3] College of Engineering Technology, University of Doha for Science and Technology, Doha P.O. Box 24449, Qatar
* Correspondence: tesau@dal.ca; Tel.: +1-(902)-893-3055

Abstract: The production of wild blueberries (*Vaccinium angustifolium*) contributes 112.2 million dollars yearly to Canada's revenue, which can be further increased by reducing harvest losses. A precise prediction of blueberry harvest losses is necessary to mitigate such losses. The performance of three machine learning (ML) algorithms was assessed to predict the wild blueberry harvest losses on the ground. The data from four commercial fields in Atlantic Canada (including Tracadie, Frank Webb, Small Scott, and Cooper fields) were utilized to achieve the goal. Wild blueberry losses (fruit loss on ground, leaf losses, blower losses) and yield were measured manually from randomly selected plots during mechanical harvesting. The plant height of wild blueberry, field slope, and fruit zone readings were collected from each of the plots. For the purpose of predicting ground loss as a function of fruit zone, plant height, fruit production, slope, leaf loss, and blower damage, three ML models i.e., support vector regression (SVR), linear regression (LR), and random forest (RF)—were used. Statistical parameters i.e., mean absolute error (*MAE*), root mean square error (*RMSE*), and coefficient of determination (R^2), were used to assess the prediction accuracy of the models. The results of the correlation matrices showed that the blueberry yield and losses (leaf loss, blower loss) had medium to strong correlations accessed based on the correlation coefficient (*r*) range 0.37–0.79. The LR model showed the foremost predictions of ground loss as compared to all the other models analyzed. Tracadie, Frank Webb, Small Scott, and Cooper had R^2 values of 0.87, 0.91, 0.91, and 0.73, respectively. Support vector regression performed comparatively better at all the fields i.e., R^2 = 0.93 (Frank Webb field), R^2 = 0.88 (Tracadie), and R^2 = 0.79 (Cooper) except Small Scott field with R^2 = 0.07. When comparing the actual and anticipated ground loss, the SVR performed best (R^2 = 0.79–0.93) as compared to the other two algorithms i.e., LR (R^2 = 0.73 to 0.92), and RF (R^2 = 0.53 to 0.89) for the three fields. The outcomes revealed that these ML algorithms can be useful in predicting ground losses during wild blueberry harvesting in the selected fields.

Keywords: machine learning algorithms; harvesting losses; wild blueberries

Citation: Khan, H.; Esau, T.J.; Farooque, A.A.; Abbas, F. Wild Blueberry Harvesting Losses Predicted with Selective Machine Learning Algorithms. *Agriculture* **2022**, *12*, 1657. https://doi.org/10.3390/agriculture12101657

Academic Editor: Xiuliang Jin

Received: 10 August 2022
Accepted: 13 September 2022
Published: 10 October 2022

1. Introduction

Native to the northern parts of North America, the wild or lowbush blueberry (Vaccinium Augustifolium Ait.) is an eternal, deciduous shrub. [1]. Canada produced 161,346 tons of harvested wild blueberries in 2020 making its production greater than 50% of the world's wild blueberries [2]. Unlike other fruits, wild blueberries grow naturally from indigenous stands on deforested lands developed for agriculture [3]. Commercial fields of wild blueberries are grown on abandoned farmland or cleared forests where domestic blueberry plants already exist [4]. The fields are clipped to the ground level in the first year (vegetative year) as part of a biennial process that primarily controls the stands and is harvested in the second year (fruiting year) [4]. The wild blueberries are small and

soft fruits with high economic value thanks to their delicious taste, rich in nutrients, and anticancer properties [5].

Wild blueberries are harvested manually or mechanically for almost 100 years. Harvesting losses occur but can be minimized with improved management practices [6]. Mechanical harvesters replaced conventional hand raking crews to increase harvest efficiency and reduce the reliance on manual labour [7].

Mechanical harvesting significantly improved the production of wild blueberries in North America since their commercialization in the early 1980s [8]. Wild blueberries are susceptible to mechanical damage due to their soft texture [9]. Reference [10] found mechanical highbush blueberry harvesters caused increased fruit bruising and harvesting losses as compared to hand raking. Variability in wild blueberry losses may be due to endogenous or exogenous factors. Yielding nature of different clones and natural soil variations in the selected fields are the intrinsic factors, while other factors like, crop management practices, terrain, harvester operation, and operator skill may include in extrinsic factors [11].

Efforts continued to improve mechanical harvesters to reduce harvesting losses. Peterson [12] redesigned an experimental highbush blueberry harvester and reported 6.9, and 8.6% harvesting losses for the redesigned experimental and commercial harvester (rotary-style). Farooque [13] reported an average blueberry yield was 8000 kg ha^{-1} in the well-maintained blueberry fields in Central Nova Scotia and above 10% of the blueberry loss was observed when mechanically harvested. They also reported that crop damage is directly proportional to the yield of harvestable fruit. Holshouser [14] evaluated that the harvesting losses could alter from 3 to 10% because of lodging in the fields of soybean. Lodging could lead to reduce picking efficiency and enhance the losses during harvesting [15]. The harvester picker bars must be in touch with the top one-third of the plant to acquire optimal soybean yield [16]. Keeping wild blueberries near open spaces will reduce fruit zone, possibly leading to more loss of fruit [13].

Traditionally, wild blueberry growers depend upon their experience and previous years data such as weather conditions, crop yield, and losses to make key determinations to enhance both the brief financial success and long-term business viability [17]. There are some commonly used methods to predict crop yield. For example, Prasad [18] used an empirical equation and its related coefficient, based on historical, meteorological and satellite data to predict wheat and rice production. They demonstrated that it is a promising technique for predicting crop productivity. Feed Forward Neural Network and Recurrent Neural Network can also be used to predict the fruit/crop yield on the basis of suitable crop parameters like TemperatureMin, TemperatureMax, humidity, wind speed and pressure [19]. Promising new technologies like machine learning (ML) have appeared more recently and can potentially aid farmers' decision-making [20].

A subtype of artificial intelligence called machine learning aims to learn from the existing data to help the growers in making informed decisions. This approach can identify patterns and correlations and uncover insights from the datasets. The models should be trained using a dataset, where the model results are expressed on the basis of experience. The predictive model is built using various aspects, and as such, the parameters of the model are decided by the use of historical information when training. In the testing stage, performance is accessed using a portion of the past records which has not yet been utilized for training. The ML models can be predictive or descriptive, depending upon the research issues and quarries. In order to learn from the gathered data and elaborate what has occurred, descriptive models are utilized whereas, estimations are made using predictive models [21]. The ML studies cover different challenges as they aim to build an effective predictive model. The selection of appropriate model is required to address the issue at hand, and the underlying platforms and models also must have the ability to manage the volume of data [22].

According to the literature's findings [23], the Random Forest (RF) model outperformed the Selection Operator regression, Least Absolute Shrinkage and ridge regression,

and extreme gradient boosting in terms of predicting maize production and Nitrogen losses. Yoosefzadeh-Najafabadi [24] compared three commonly used ML models namely RF, multi-layer perception (MLP), and support vector machine (SVM) to predict soybean yield. Their results revealed that the RF had the highest prediction accuracy in predicting soybean yield as compared to the other two models they tested. Esfandiarpour-Boroujeni [25] estimated the apricot yield with high accuracy (R^2 = 0.81) using support vector regression (SVR). Abbas [26] predicted potato yield using four ML models namely LR, k-nearest neighbour, elastic net, and SVR concluded that all the algorithms worked very well in explaining the tuber yield having R^2 = 0.70, 0.65, 0.64, and 0.72 respectively.

The literature review has shown that various ML models have been used for the prediction of crop yield and loss. However, limited work has been done using ML models to predict the wild blueberry harvesting fruit losses. There is a need to investigate harvesting patterns and pinpoint losses since wild blueberry growers experience significant harvesting losses as a result of modified growing circumstances brought on by novel management techniques [27]. Prediction of harvesting losses would help farmers in decision-making so that they can develop their harvesting strategies to overcome the predicted losses by increasing the fruit yield. The goal of this study was to predict wild blueberry ground losses during harvesting using ML models.

2. Methodology

2.1. Data Sites

Data about blueberry mechanical harvesting yield losses and the related factors contributing to the yield losses were obtained from four wild blueberry field studies conducted in Nova Scotia. The selected sites were in commercial wild blueberry fields including Frank Webb (45.404733° N, 63.669376° W), Tracadie (47.511270° N, 65.138270° W), Cooper (45.480573° N, 63.573471° W), and Small Scott (45.600641° N, 63.086512° W) having field areas of 2.57, 1.6, 3.2, and 1.9 ha, respectively. While Cooper field and the Small Scott field both were in the year of vegetation in 2010 and their fruit-bearing year in 2011, Frank Webb and Tracadie were in their year of vegetation in 2011 and their fruit-bearing year in 2012. To replicate early and late harvesting, the chosen fields were harvested every year from early August to early September using a mechanical blueberry harvester (Doug Bragg Enterprises Ltd., Collingwood, NS, Canada). Figure 1 displays the geolocation of the chosen fields. The chosen fields had undergone biennial trimming by mowing as well as traditional farming management techniques, and they had been commercially managed for the previous ten years (fertilization, pruning, weed, disease, and pollination).

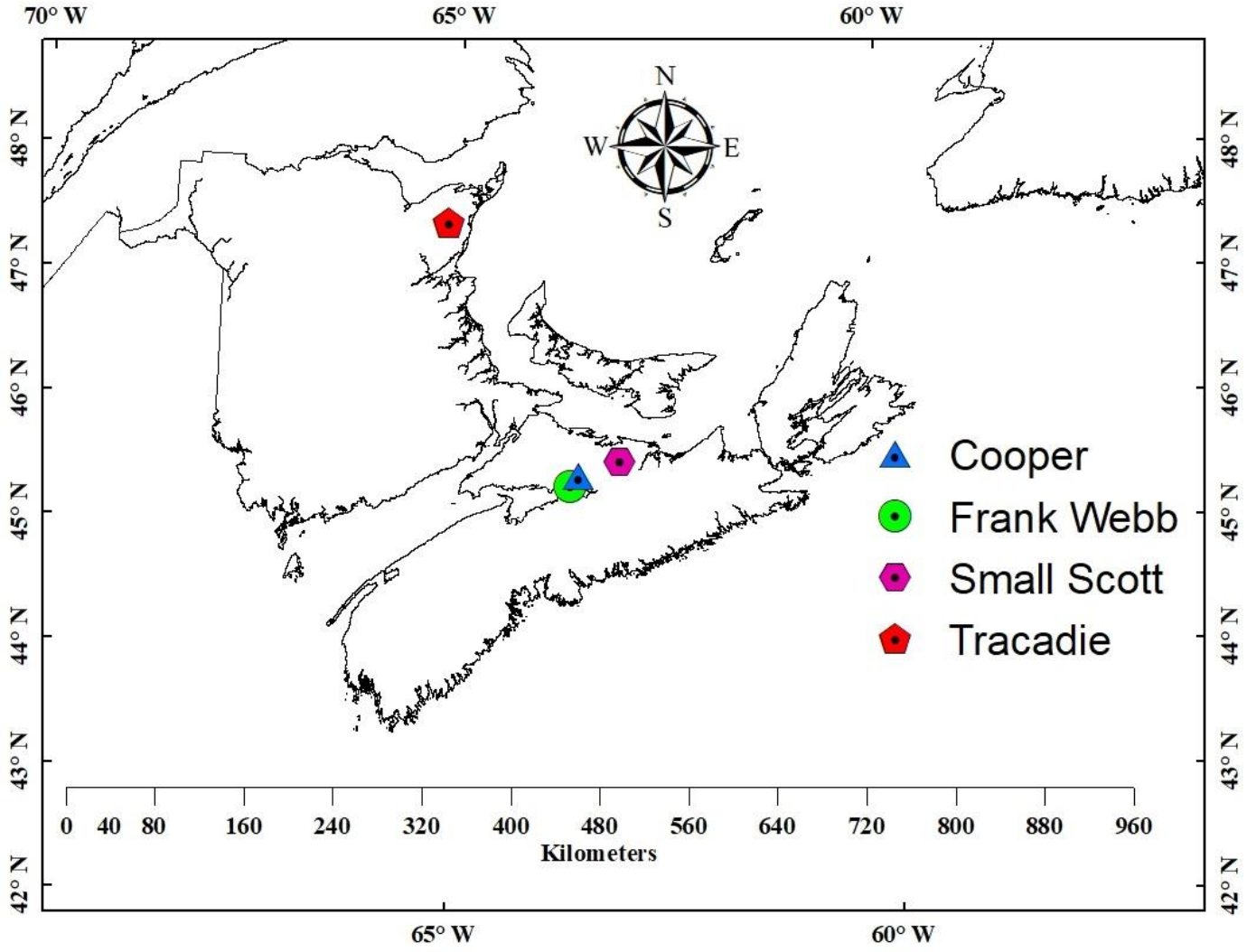


Figure 1. Selected blueberry fields in two Atlantic provinces (New Brunswick and Nova Scotia).

2.2. Data Collection and Analysis

For data collection to understand the harvesting losses, eighty-two plots of 0.91 × 3 m dimension (identical to harvester head's width) were flagged arbitrarily in the Frank Webb field, Cooper field, and Small Scott field, and one hundred and nine plots were flagged in the Tracadie field. Each plot included a 0.3 m buffer around it to prevent inaccuracy during the data collection. A John Deere tractor (62.5 kW) was equipped with a solitary wild blueberry harvester. (Moline, Ill., Grand Detour, IL, United States). The harvester was operated in the fields at a ground speed of 1.6 km h^{-1} and 28 rpm. At the beginning of all the plots, the harvester's head was put down for harvesting and then raised at the end point of the plot. The belt of the harvester conveyor was connected to a bucket to collect the blueberries from each plot. Three losses including blower, ground, and leaf losses were considered. The blower damage was retrieved by mounting a collection bucket below the harvester blower fans that was emptied after each plot. The dropped berries were hand-picked from each plot to calculate the ground loss. For the leaf loss, the leaves and debris were separated from the collected good berries, placed in labelled Ziploc®bags, and measured to calculate the weight of yield and fruit loss.

For average plant height per field, five plants were selected in each plot to measure their height. Readings of plant height were measured using a measuring tape and then averaged for each plot. The zone from the top to the bottom of the cluster of fruits on blueberry plants is indicated as the fruit zone. The purpose of fruit zone reading was

to help the operator in adjusting the harvester's head height from the ground to pick blueberries effectively. Slope measurements (five at each plot) were recorded by hand using a Craftsman SmartTool Plus digital level (Sears Holdings Corporation, Hoffman Estates, IL, USA) and then averaged to get a characteristic slope for the selected fields from the slope values of each plot. Fruit zone and plant height ranged from 7.4 to 34.6 cm and 10.6 to 39.0 cm respectively and both were moderately variable. The slope was highly variable with a range from 0.2 to 23.7 degrees within all the selected fields. The attributes of the harvesting plots are given in Table 1.

Table 1. Descriptive statistics of harvesting plot attributes.

Field	No. of Plots	Plot Attribute	Mean ± SD	Minimum	Maximum	CV (%)
Frank Webb	82	Plant Height (cm)	22.4 ± 3.66	13.0	31.8	16.3
Tracadie	109		23.7 ± 3.83	19.0	39.0	15.0
Cooper	82		23.6 ± 4.06	10.6	32.8	17.2
Small Scott	82		24.0 ± 3.63	13.0	34.0	15.8
Frank Webb	82	Fruit Zone (cm)	17.6 ± 3.43	11.0	24.8	19.6
Tracadie	109		22.8 ± 4.00	11.2	34.6	17.5
Cooper	82		19.4 ± 3.57	7.80	25.3	18.4
Small Scott	82		19.1 ± 3.62	7.40	31.0	19.0
Frank Webb	82	Slope (degree)	7.86 ± 5.16	0.73	21.7	65.7
Tracadie	109		2.48 ± 1.35	0.47	6.57	54.4
Cooper	82		7.47 ± 4.40	0.50	19.5	58.9
Small Scott	82		7.04 ± 4.48	0.20	23.7	63.6

CV = coefficient of variation; SD = standard deviation.

The primary sign of the variability is the coefficient of variation (CV) in descriptive statistics. A CV less than 15% shows the least variability of parameters; the CV between 15 to 35% indicates that the parameter is moderately variable and the CV greater than 35% describes that the parameter is highly variable [28]. The ground loss varied from 3.4 to 1847 kg/ha with CV > 35% across all the fields. Relationships of all variables were assessed using Pearson correlation coefficients. The values of correlation coefficients ($r \leq 0.35$) normally stand for weak correlations, 0.36–0.67 for medium correlations, and 0.68 to 0.90 for strong correlations, and $r \geq 0.90$ shows significantly high correlations [29]. The selected ML models were trained using datasets (plant size, fruit area, slope, blueberry yield, leaf loss, and the blower loss). Because the model was constructed utilizing a variety of attributes, its parameters were established during the training stage using data from prior years (2011 and 2012). During the testing stage, the fraction of the data which wasn't utilized for training was employed for performance assessment. [22].

2.3. Machine Learning Models

Machine learning can direct patterns and correlations and uncover insights from the datasets. 80% of the data was utilized in training and 20% in testing. Machine learning studies consist of different challenges like inaccessible data, data security, and time-consuming implementation when building a well-functioning predictive model. It is difficult to choose the correct models to solve the issue at hand, and moreover, the models and the fundamental platforms have to handle a big volume of data [22].

2.3.1. Linear Regression

Linear Regression (LR) models are understandable but incredibly powerful. Linear regression gives an impact of each predictor variable on the response variable [30]. They reported that supervised learning is a method used in LR. It can be applied to predict continuous variables. Linear regression in ML uses data to learn by reducing the loss commonly known as the mean square error (*MSE*) or root mean square error (*RMSE*) by using models, for example, gradient descent. Based on the type of data, the gradient

descent model works at minimal loss functions, increasing the LR model's ability to predict outcomes accurately. [30]. Linear regression is described by the equation below:

$$y = a + bx$$

where a = constant intercept, b = slope of a regression line.

The loss function (J) assists in evaluating the values of coefficients (a, b) by reducing the inaccuracy in between the real values and the anticipated values. It can be explained by the equation below:

$$\text{Minimize } J = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$$

where, $\hat{y}_i$ = predicted value, y_i = actual value.

2.3.2. Support Vector Regression

Support vector regression is a supervised learning model that may be utilized for both regression and classification purposes [31,32]. Support vector regression can be linear or non-linear using respective kernel functions. The well-known kernels are linear, radial basis function, sigmoidal, polynomial. The productivity of the SVR very much realize on the selection of the kernel. Linear kernel is used in the SVR for linear regression while using as appropriate nonlinear kernel makes it nonlinear [33]. By including the hyperplane and widening the gap between the anticipated and real values, SVR has always aimed to reduce inaccuracy. The SVR may perform better when applied to information that is imbalanced regarding the binary outcome because they are created utilizing only the support vectors [34]. On the other hand, it has some drawbacks, for example, the user must choose the SVR's kernels for nonlinear scenarios. The kernel and any related hyperparameters that the kernel requires should be specifically picked; a bad kernel selection might impair the performance of the model. [35]. Linear SVR was used for this study based on the results obtained from LR. Linear SVR can be defined by the formula given below:

$$y = \sum_{i=1}^{n}(a_i - a_i^*)k(a_i, x) + b$$

where, a and x = supplementary hyperplanes in conjunction with the regression line.

2.3.3. Random Forest

The RF model is a form of ensemble approach which generates forecasts by aggregating forecasts from many different base models. The RF model has had outstanding luck as a particular regression and classification tool since its inception by [36]. The bootstrap aggregating technique used by the RF model, also known as bagging, lowers the variability of a quantitative learning approach. [37]. In summary, different bootstrapped specimens out of the training information are collected, and trees are built using these samples. A democratic decision is made for each tree's anticipated class, and an average forecast is then returned. The overall prediction power of the model is potentially increased by this method. In addition, an estimate of out-of-bag error, which is a reliable estimation of the test error, is possible with bootstrap aggregating [36]. Both regression and classification problems can be solved by the RF model, which makes it a diverse model that is extensively used by engineers. In addition to prediction accuracy, A wide range of industries, especially the share market, banks, pharmacology, patient healthcare management, and physiology, frequently use RF as a tool [38]. While using the RF to solve regression problems, *MSE* has been used to know data branches from each node. The following equation is used to find *MSE*:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where, N = total number of points, f_i = output of the model, and y_i = true value of data point i.

The RF model can provide the estimation of the importance of the variables by comparing the changes of *MSE* when a specific variable is randomly altered, and other variables are kept unchanged.

2.3.4. Hyperparameters Tunning

The data samples were split up into 20% and 80% sets for the testing and training of the data sets, respectively. It is important to evaluate various hyperparameters for varied datasets, because all hyperparameters behave uniquely for the different kinds of datasets [26]. That is why the testing procedure was performed by adopting the hit and trial method based on which the best combination of values was selected which gave the highest R^2, mean absolute error (*MAE*) and *RMSE*. Following this procedure, the hyperparameters which are displayed in Table 2 were utilized to train the selected ML models. The hit and trial method was adopted to determine the range of hyperparameters. Support vector regression was tested by optimizing the regularization parameter (C) value from 50–200 and it performed well at C = 150. Similarly, the Epsilon value was optimized from 0.1–1.0, and it performed well at 0.2. In case of RF, seven hyperparameters were tested at maximum depth = 10–60, random state = 5–75, min samples leaf = 1–20, verbose = 0.1–10 and it showed the best results at maximum depth = 35, random state = 30, min samples leaf = 3, and verbose = 2.

Table 2. Hyperparameters tuning of machine learning models.

Algorithms	Hyperparameters	
Linear Regression	Intercept calculation (fit intercept)	TRUE
	Data normalization	FALSE
	Number of iterations (n jobs)	None
	True X copying	TRUE
Support Vector Regression	Defining algorithms (kernel)	Linear
	Regularization parameter (C)	150.0
	Penalty association (Epsilon)	0.2
Random Forest	Maximum depth	35.0
	Random state	30.0
	Min samples split	6.0
	Min samples leaf	3.0
	Max features	8.0
	Max leaf nodes	None
	Verbose	2.0

2.3.5. Model Evaluation Criteria

References [39,40] used three statistical parameters, R^2, *RMSE*, and *MAE*, which were utilized to evaluate LR, SVR, and RF models. R^2 assesses that how well a model explains or predicts the outcomes. Its value lies from 0.0 to 1.0 range. A value closure to 1 represents the model's excellent efficiency.

$$R^2 = \sqrt{\frac{\sum_{i=1}^{N}(yi - \overline{y})^2 - \sum_{i=1}^{N}(yi - \hat{y}i)^2}{\sum_{1=1}^{N}(yi - \overline{y})^2}}$$

The amount of error in a measurement is called absolute error and an average of all those absolute errors is known as *MAE*.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)$$

The difference between both the true and anticipated values are measured using *RMSE*. The effectiveness of the model is indicated by a reduced *RMSE* value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(yi - \hat{y}i)^2}{N}}$$

where, yi = actual value present at the ith time, $\hat{y}i$ = estimated value at the ith time, $\overline{y}$ = mean value of yi, i has a value range of 1 to N, and N = number of values.

3. Results and Discussion

3.1. Descriptive Statistics

Table 3 displays the findings of the descriptive statistics for the chosen parameters. Fruit yield was highly variable within all the selected fields with values varying from 253 to 17,968 kg/ha. The ground loss were highly variable and occurred due to many factors which include pre-harvest berry drop. Leaf loss showed a high variability ranging from 0 to 575 kg/ha. Blower loss also showed high variability (0–529 kg/ha) except only moderate variability for at the Tracadie site (21.2–129 kg/ha).

Table 3. Descriptive statistics of chosen parameters.

Field	Variable	Mean ± SD	Minimum	Maximum	CV (%)
Frank Webb	Fruit Yield (kg/ha)	8136 ± 2914	2218	17968	35.8
Tracadie		5572 ± 2102	1690	13574	37.7
Cooper		3705 ± 2014	305	9914	54.4
Small Scott		2618 ± 1570	253	7635	60.0
Frank Webb	Ground Loss (kg/ha)	1072 ± 386	132	1847	36.0
Tracadie		580 ± 217	148	1056	37.4
Cooper		291 ± 186	19.6	891	63.9
Small Scott		165 ± 127	3.40	708	77.0
Frank Webb	Leaf Loss (kg/ha)	244 ± 116	42.9	575	47.4
Tracadie		88.2 ± 34.4	23.8	320	39.5
Cooper		83.9 ± 77.4	4.90	343	92.6
Small Scott		39.7 ± 62.7	0	299	158
Frank Webb	Blower Loss (kg/ha)	142 ± 90.6	31.5	529	63.8
Tracadie		67.8 ± 20.4	21.2	129	30.0
Cooper		43.5 ± 39.1	4.90	225	89.8
Small Scott		22.2 ± 33.0	0	220	149

CV = coefficient of variation; SD = standard deviation.

3.2. Correlation Analysis

In order to identify the relationships in between ground losses and other input variables, correlation matrices were established. The Pearson correlation's results have been shown in Figure 2. In the Frank Webb field, there were strong significant, and positive correlations between ground loss and fruit yield (r = 0.78), and leaf loss (r = 0.79). Farooque [41] reported that the fruit losses on the ground enhanced with an increment in the

blueberry yield during the harvesting. There was a moderate correlation between ground loss and blower loss (r = 0.62). The ground loss was negatively correlated with plant height (r = −0.28) and fruit zone (r = −0.06) which means that ground loss decreased while plant height and fruit zone increased and vice versa. It has also been reported by [13] that the ground loss was inversely proportional to plant height (r = −0.21) and fruit zone (r = −0.07). The slope had a positive correlation with ground loss (r = 0.04). In the Cooper field, the ground loss had a moderate positive correlation with fruit yield (r = 0.47). It was due to the topography of the field and the size of berries. The remaining variables had a weak correlation i.e., $r \leq 0.35$. In the Small Scott field, ground loss and fruit yield were positively correlated (r = 0.59). Leaf loss, blower loss, and slope also had a positive correlation with ground loss i.e., r = 0.33, 0.14, and 0.37, respectively. In the Tracadie field, a significant correlation between ground loss and fruit yield was observed (r = 0.73). Leaf loss, blower loss, and slope were weakly correlated with the ground loss which means that $r \leq 0.35$ for these variables. [13] concluded that ground loss had a significant correlation with fruit yield (r = 0.78) but it had a weak correlation with blower loss (r = 0.15) and slope (r = 0.16). Plant height and fruit zone represented a reverse relationship with the ground loss.

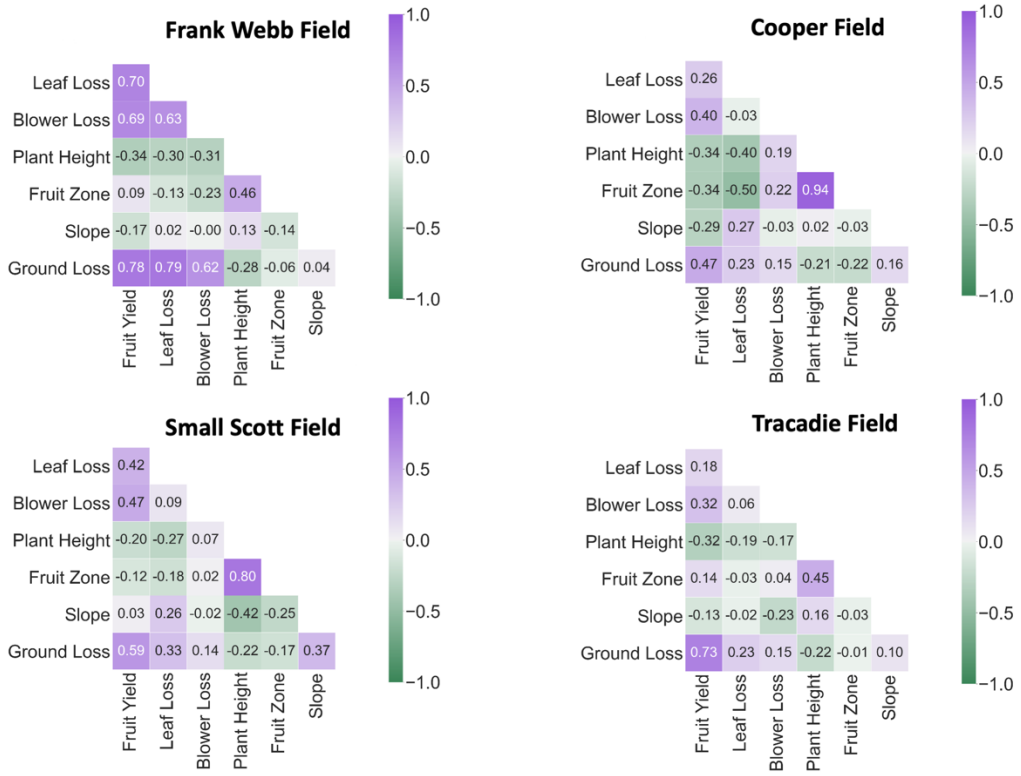


Figure 2. Pearson correlation analysis of selected variables for four fields in which all possible relationships within all variables are presented.

3.3. Evaluation of Machine Learning Algorithms

The outcomes of the model assessment have been given in the Table 4. SVR had a higher R^2 (0.93) for Frank Webb field; LR recorded R^2 = 0.91 whereas, the lowest R^2 (0.53) was recorded for RF in this field. The ranges of *MAE* and *RMSE* were 2.35–2.49 and 2.96–3 kg/ha respectively for all the algorithms in this field. In the Tracadie field, high R^2 was recorded for LR and SVR which were 0.87 and 0.88 respectively, whereas, RF had R^2 = 0.78. The values of *MAE* and *RMSE* for this field ranged from 10.74–34.32 and 13.08–45.15 kg/ha, respectively. In Cooper field higher R^2 (0.89) was observed for RF whereas, for LR and SVR, the values of R^2 were 0.73 and 0.79, respectively. Lowest values of *MAE* and *RMSE* were calculated for SVR i.e., 0.1 and 0.15, respectively in this field. In the Small Scott field higher R^2 value was recorded for LR (0.91), and the lowest SVR and RF were recorded at

0.07 and 0.18 for this field, respectively. The highest *MAE* and *RMSE* were observed for RF i.e., 53.76 and 103, respectively. The findings revealed that the SVR and LR performed very well in predicting the berry losses (Table 4). Wang [42] compared the performance of the RF algorithm with SVR and artificial neural network (ANN) to remotely estimate the wheat biomass and reported that RF (R^2 = 0.79) and SVR (R^2 = 0.62) showed good results as compared to ANN (R^2 = 0.3). Gandhi [43] used different machine learning techniques and reported that SVR performed very well for the prediction of rice crop yield under different climatic scenarios. Palanivel [44] used diverse machine learning techniques such as LR, ANN, and backpropagation methods to predict the crop yield. In order to develop a prediction model for fruit yield, Obsie [45] selected four ML models, that are boosted decision trees, multiple linear regression, extreme gradient boosting, and RF, and concluded that RF was the second-most successful algorithm, accompanied by the Boosted Decision Tree algorithm with R^2 = 0.90.

Table 4. Comparison of algorithms (linear regression, support vector regression, and random forest) within fields.

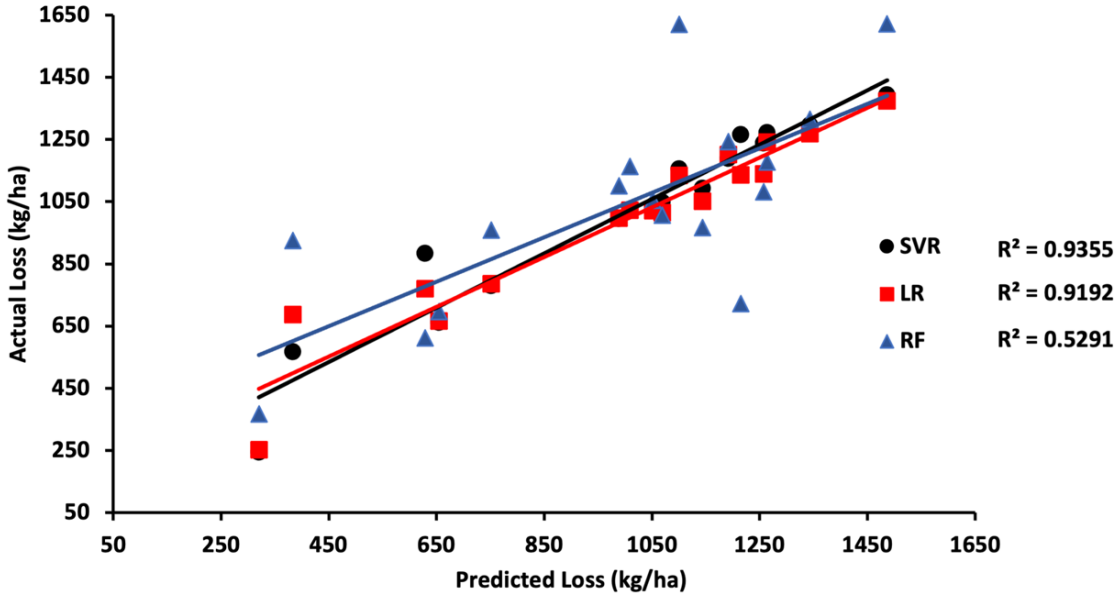
Field	Algorithm	MAE (kg/ha)	RMSE (kg/ha)	R^2
Frank Webb	Linear Regression	2.35	2.96	0.91
	Support Vector Regression	2.46	3.22	0.93
	Random Forest	2.49	3.00	0.53
Tracadie	Linear Regression	10.7	13.1	0.87
	Support Vector Regression	10.6	12.8	0.88
	Random Forest	34.3	45.2	0.78
Cooper	Linear Regression	1.95	3.01	0.73
	Support Vector Regression	0.10	0.15	0.79
	Random Forest	53.7	103	0.89
Small Scott	Linear Regression	1.95	3.01	0.91
	Support Vector Regression	0.14	0.18	0.07
	Random Forest	53.8	103	0.18

RMSE, root means square error; *MAE*, mean absolute error.

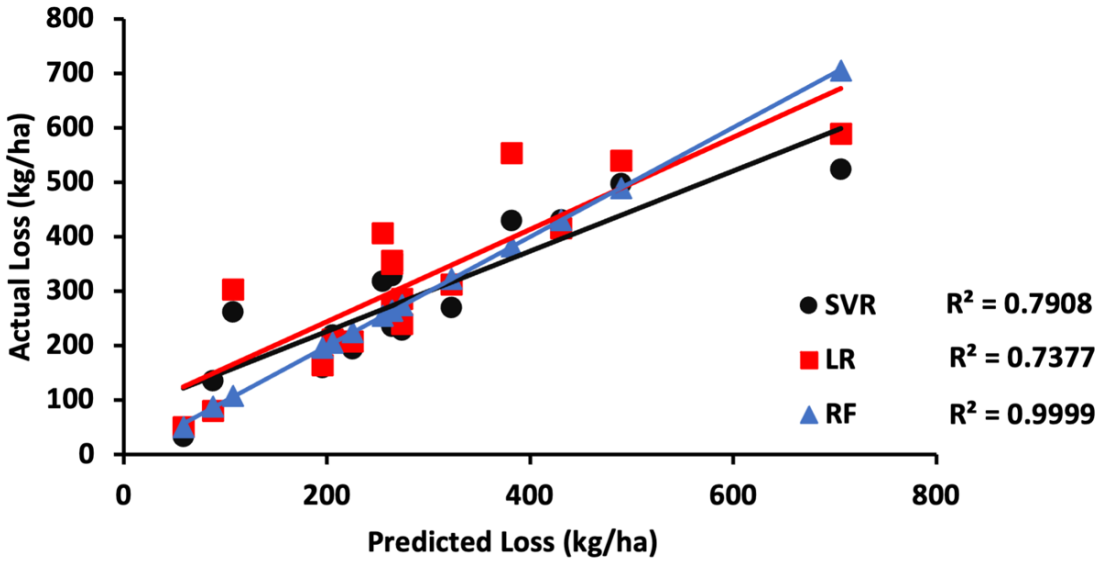
3.4. Comparison of Actual and Predicted Ground Losses

Outputs of algorithms were compared to evaluate which algorithm performed better in predicting the ground losses from plant size, fruit area, topography, blueberry yield, leaf damage, and the blower loss as shown in Figure 3. In the Frank Webb field, SVR (R^2 = 0.94) performed better as compared to LR (R^2 = 0.91) in predicting the ground losses. In this field, RF did not perform well in predicting the losses (R^2 = 0.53). Whereas, in the Cooper field, RF had the highest value (R^2 = 0.99) which means RF performed very well in predicting the ground losses, while SVR and LR had (R^2 = 0.79) and (R^2 = 0.74) respectively. LR was found to be the best performer in predicting the ground losses in Small Scott (R^2 = 0.91) and Tracadie (R^2 = 0.89). Whereas SVR (R^2 = 0.88) and RF (R^2 = 0.78) were also good in predicting the ground losses for the Tracadie field. In comparison, SVR and RF performed better in three fields except for the Small Scott field LR performed well in all the fields. The logic behind poor performance of RF and SVR in the Small Scott field could be the result of influencing factors such as climate, soil, etc. which may influence yield. Therefore, some unknown factors which were not included in this study may have influence on yield which reduced modelling accuracy at the Small Scott site. Different models performed well in different studies like [46] used machine learning algorithms namely SVR, RF, and deep neural networks for the autumn crop yield prediction. The results showed that SVR and RF performed very well in predicting the yield having R^2 = 0.92 and R^2 = 0.90. All the models performed differently in varying fields due to the type of data. They performed well,

especially for linear data and our data is point-based or discrete data which is not linear. So, the performance of the models depends on the correlation between input parameters (slope, plant height, blower loss, fruit zone, and leaf loss) and the output data (ground loss) which is different for each of the fields. The models performed differently in each fields because the productivity of models relies on the nature of the specific input data.

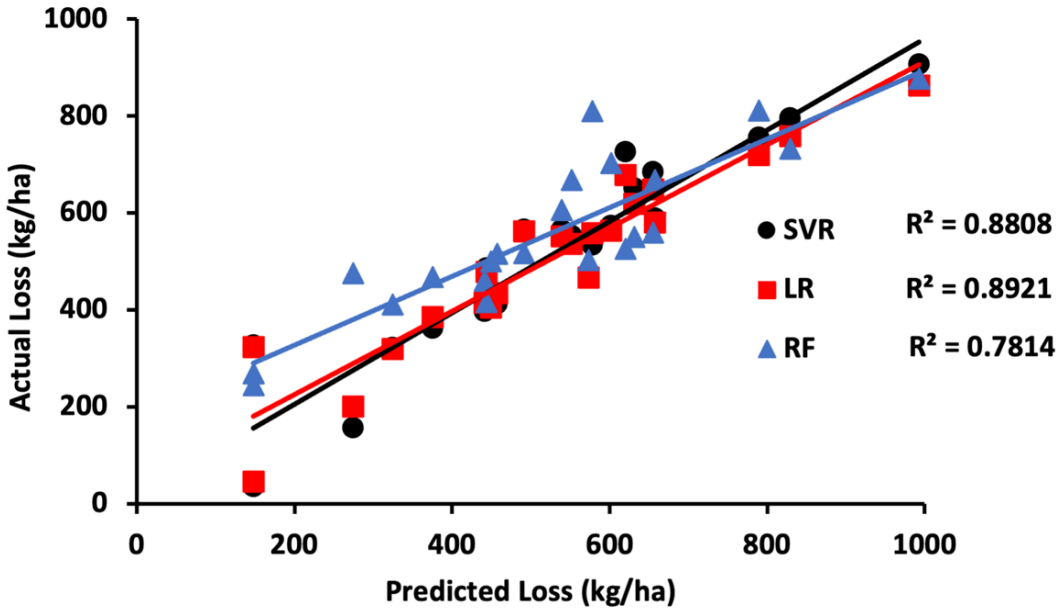


(**a**) Frank Webb

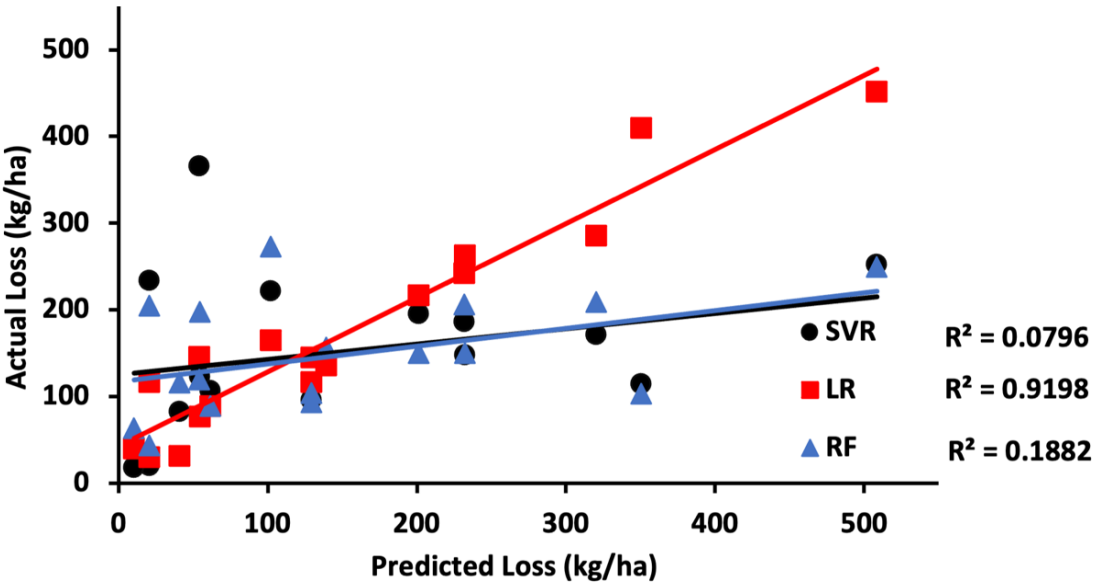


(**b**) Cooper

Figure 3. *Cont.*



(**c**) Tracadie



(**d**) Small Scott

Figure 3. Comparison of actual and estimated ground loss within the selected fields.

3.5. Comparison of Machine Learning Algorithms

Three ML models were utilized in this study to find the ground losses. The comparison of these algorithms showed that the LR and SVR performed comparatively well for all the fields as shown in Figure 4. The LR performed better because it utilizes the data to learn by reducing loss like *MAE* and *RMSE* [29]. SVR may perform better than other algorithms due to its use of a stronger optimization method for a wide range of variables [47]. Pan [48] established quantitative structure-function relationship algorithms for forecasting the auto-ignition temperatures of organic substances using a support vector. Investigated and contrasted the calibration and predictive power of the SVR with the other two widely used techniques, back-propagation neural network and LR. Outcomes revealed that the support vector performed better as compared to the backpropagation and MLR.

Additionally, it demonstrated improved generalization capabilities for the support vector and demonstrated that it is a powerful resource. The result of this research also highlights the superior productivity of LR and SVR in comparison to RF because of their improved optimization methods for a large number of parameters [46]. Support vector regression gives the supplemental functionality of kernel, which increases the productivity of the model by understanding the nature of attributes [49]. Linear regression performance was best in all the fields and SVR performance was better for three of the four fields. Whereas RF performed well for only two fields. On the basis of this study's findings, LR and SVR models are suggested to predict the ground losses in the selected blueberry fields.

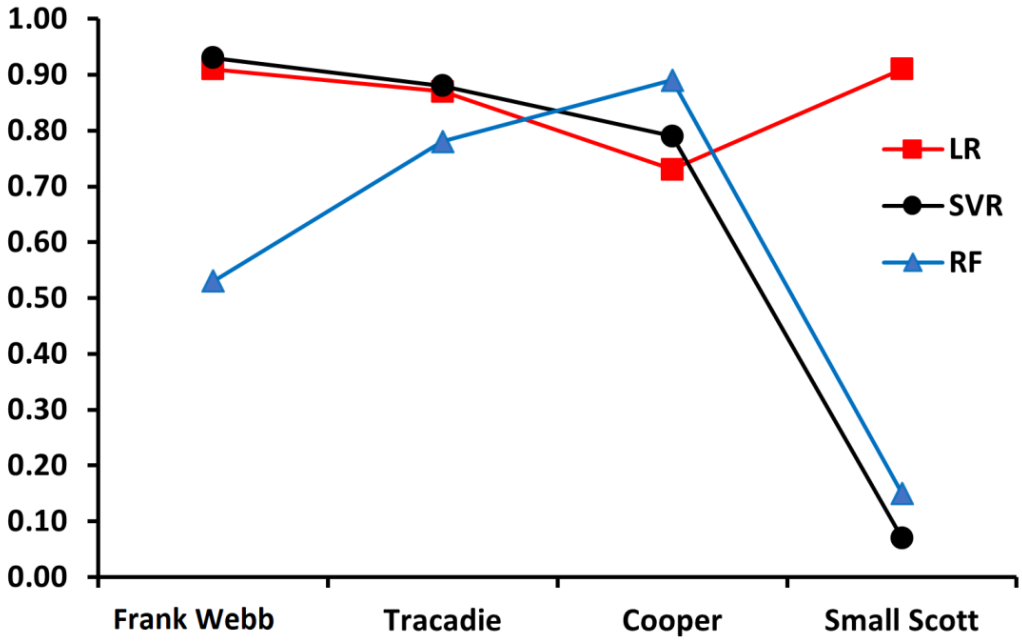


Figure 4. Comparison of machine learning algorithms within selected fields.

4. Conclusions

In this study, the losses on the ground have been predicted during the harvesting of blueberry using ML algorithms and the best algorithms have been proposed which can be used to predict the fruit losses on the ground. Four blueberry fields were selected, and a randomized experiment was conducted in each field. Eighty-two plots were setup in three fields and one hundred and nine plots were made in the fourth field. Berry losses and fruit yield were measured from each plot. The values of fruit zone, plant height, and topography were also noted from all the plots within the selected fields. Three ML algorithms namely LR, SVR, and RF were used to predict ground losses. Modeling techniques were used to access the prediction of ground losses. Findings of correlation investigation indicated that the blueberry yield and the losses (leaf loss, blower loss) had moderate to high correlations with the ground loss with r ranging from 0.37–0.79. LR model performed best as compared to the other models for Frank Webb, Tracadie, Cooper, and Small Scott with R^2 = 0.91, 0.87, 0.73, and 0.91, respectively. With the exception of Small Scott (R^2 = 0.07), the SVR model also outperformed the competition for the Frank Webb (R^2 = 0.93), the Tracadie (R^2 = 0.88) and the Cooper (R^2 = 0.79). When actual and anticipated ground losses are compared, the LR model performed best with R^2 ranging from 0.73–0.92 within all selected fields. SVR also performed well with R^2 ranging from 0.79 to 0.93 for three fields. The results showed that these ML algorithms could be used to predict blueberry losses on the ground. These results will further help in optimizing the harvesting techniques.

Author Contributions: Formulation: T.J.E. and A.A.F., material and methods: H.K. and A.A.F., software: H.K., validation: H.K. and A.A.F., investigation: H.K. and A.A.F., proper analysis: H.K. and F.A., resources, T.J.E., data formulation: H.K. and A.A.F., writing—primary draft: H.K. and F.A., writing—review and modification: T.J.E., A.A.F. and F.A., visualization: H.K., project administration: T.J.E., supervision, T.J.E., funding acquisition: T.J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the following grant sources: Doug Bragg Enterprises and Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Development (CRD) Grants Program, and New Brunswick Canadian Agricultural Partnership (CAP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Doug Bragg Enterprises Ltd., NSERC, and New Brunswick CAP for their financial support in completing this project. The authors would also like to thank Doug Wyllie (farm manager, Bragg Lumber Company) for the provision of commercial fields for data collection.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vander Kloet, S.P. *The Genus Vaccinium in North America*; Research Branch, Agriculture Canada: Ottawa, ON, Canada, 1988.
2. Statistics Canada. Table 32-10-0364-01. Area, Production and Farm Gate Value of Marketed Fruits. 2020. Available online: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210036401 (accessed on 16 September 2022). [CrossRef]
3. Hall, I.V.; Aalders, L.E.; Nickerson, N.L.; Kloet, S.P.V. The biological flora of Canada. 1. Vacinnium angustifolium Ait., Sweet lowbush blueberry. *Can. Field Nat.* **1979**, *93*, 415–430.
4. Agriculture and Agri Food Canada. Crop Profile for Wild Blueberry in Canada. 2005. Available online: https://publications.gc.ca/site/fra/9.689941/publication.html (accessed on 15 September 2022).
5. Baby, B.; Antony, P.; Vijayan, R. Antioxidant and anticancer properties of berries. *Crit. Rev. Food Sci. Nutr.* **2017**, *15*, 2491–2507. [CrossRef] [PubMed]
6. Yarborough, D.E.; Hergeri, G.B. Mechanical harvesting of berry crops. *Hortic. Rev.* **2010**, *16*, 255–282.
7. Yarborough, D.E. Progress towards the Development of a Mechanical Harvester for Wild Blueberries. Fact Sheet No. 226. University of Maine Cooperative Extension, 1992. Available online: http://umaine.edu/blueberries/factsheets/production (accessed on 21 June 2022).
8. Hall, I.V.; Craig, D.L.; Lawrence, R.A. A comparison of hand raking and mechanical harvesting of lowbush blueberries. *Can. J. Plant Sci.* **1983**, *63*, 951–954. [CrossRef]
9. Fan, S.; Li, C.; Huang, W.; Chen, L. Detection of blueberry internal bruising over time using NIR hyperspectral reflectance imaging with optimum wavelengths. *Postharvest Biol. Technol.* **2017**, *134*, 55–66. [CrossRef]
10. Peterson, D.L.; Brown, G.K. Mechanical harvester for fresh market quality blueberries. *Trans. ASAE* **1996**, *39*, 823–827. [CrossRef]
11. Hepler, P.R.; Yarborough, D.E. Natural variability in yield of lowbush blueberries. *HortScience* **1991**, *26*, 245–246. [CrossRef]
12. Peterson, D.L.; Wolford, S.D.; Timm, E.J.; Takeda, F. Fresh market quality blueberry harvester. *Trans. ASAE* **1997**, *40*, 535–540. [CrossRef]
13. Farooque, A.A.; Zaman, Q.U.; Groulx, D.; Schumann, A.W.; Yarborough, D.E.; Nguyen-Quang, T. Effect of ground speed and header revolutions on the picking efficiency of a commercial wild blueberry harvester. *Appl. Eng. Agric.* **2014**, *30*, 535–546.
14. Holshouser, D. Virginia Soybean Update. Available online: https://blogs.ext.vt.edu/soybean-update/ (accessed on 23 February 2022).
15. Woods, S.J.; Searingin, M.L. Influence of simulated early lodging upon soybean seed yield and its components 1. *Agron. J.* **1977**, *69*, 239–242. [CrossRef]
16. Huitink, G. Harvesting Soybeans. In *Arkansas Soybean Handbook*; University of Arkansas: Fayetteville, AR, USA, 2000; pp. 1–12.
17. Arbuckle, J.G.; Rosman, H. Iowa Farmers' Nitrogen Management Practices and Perspectives. 2014. Available online: https://core.ac.uk/download/pdf/38921821.pdf (accessed on 26 July 2022).
18. Prasad, A.K.; Singh, R.P.; Tare, V.; Kafatos, M. Use of vegetation index and meteorological parameters for the prediction of crop yield in India. *Int. J. Remote Sens.* **2007**, *28*, 5207–5235. [CrossRef]
19. Sivanandhini, P.; Prakash, J. Crop yield prediction analysis using feed forward and recurrent neural network. *Int. J. Innov. Sci. Res. Technol.* **2020**, *5*, 1092–1096.
20. González Sánchez, A.; Frausto Solís, J.; Ojeda Bustamante, W. Predictive ability of machine learning methods for massive crop yield prediction. *Span. J. Agric. Res.* **2014**, *12*, 313–328. [CrossRef]

21. Alpaydin, E. *Introduction to Machine Learning*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2010; pp. 486–489.
22. Van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
23. Shahhosseini, M.; Martinez-Feria, R.A.; Hu, G.; Archontoulis, S.V. Maize yield and nitrate loss prediction with machine learning algorithms. *Environ. Res. Lett.* **2019**, *14*, 124026. [CrossRef]
24. Yoosefzadeh-Najafabadi, M.; Earl, H.J.; Tulpan, D.; Sulik, J.; Eskandari, M. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield from Hyperspectral Reflectance in Soybean. *Front. Plant Sci.* **2021**, *11*, 624273. [CrossRef]
25. Esfandiarpour-Boroujeni, I.; Karimi, E.; Shirani, H.; Esmaeilizadeh, M.; Mosleh, Z. Yield prediction of apricot using a hybrid particle swarm optimization-imperialist competitive algorithm-support vector regression (PSO-ICA-SVR) method. *Sci. Hortic.* **2019**, *257*, 108756. [CrossRef]
26. Abbas, F.; Afzaal, H.; Farooque, A.A.; Tang, S. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy* **2020**, *10*, 1046. [CrossRef]
27. Farooque, A.A.; Zaman, Q.U.; Nguyen-Quang, T.; Groulx, D.; Shumann, A.W.; Chang, Y.K. Development of predictive model for wild blueberry harvester fruit losses during harvesting using artificial neural network. *Appl. Eng. Agric.* **2016**, *32*, 725–738.
28. Wilding, L.G. Spatial variability: Its documentation, accommodation and implication to soil surveys. In *Soil Spatial Variability*; Proceedings of a Workshop of ISSS and SSA, PUDOC: Las Vegas, NV, USA, 1985; pp. 166–187.
29. Taylor, R. Interpretation of correlation coefficient: A basic review. *JDMS* **1990**, *6*, 35–39. [CrossRef]
30. Ray, S. A Quick Review of Machine Learning Algorithms. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 35–39.
31. Smola, A.J.; Scholkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
32. Gunn, S.R. Support Vector Machines for Classification and Regression. *ISIS Tech. Rep.* **1998**, *14*, 5–16.
33. Lima, A.R.; Cannon, A.J.; Hsieh, W.W. Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy. *Comput. Geosci.* **2013**, *50*, 136–144. [CrossRef]
34. Attewell, P.; Monaghan, D.B.; Kwong., D. *Data Mining for the Social Sciences: An Introduction*; University of California Press: Oakland, CA, USA, 2015.
35. Horváth, G. Neural Networks in Measurement System. In *Advances in Learning Theory: Methods, Models and Applications*; Suykens, J.A.K., Horváth, G., Basu, S., Micchelli, C., Vandewalle, J., Eds.; NATO Science Series III: Computer & Systems Sciences; IOS Press: Amsterdam, The Netherlands, 2003; Volume 190, pp. 375–396.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. In *Springer Series in Statistics*; Springer: New York, NY, USA, 2001.
38. Fife, D.A.; D'Onofrio, J. Common, uncommon, and novel applications of random forest in psychological research. *Behav. Res. Methods* **2022**, 1–20. [CrossRef]
39. Nguyen, V.V.; Pham, B.T.; Vu, B.T.; Prakash, I.; Jha, S.; Shahabi, H.; Shirzadi, A.; Ba, D.N.; Kumar, R.; Chatterjee, J.M.; et al. Hybrid Machine Learning Approaches for Landslide Susceptibility Modeling. *Forests* **2019**, *10*, 157. [CrossRef]
40. Kamruzzaman, M.; Makino, Y.; Oshita, S. Rapid and non-destructive detection of chicken adulteration in minced beef using visible near-infrared hyperspectral imaging and machine learning. *J. Food Eng.* **2016**, *170*, 8–15. [CrossRef]
41. Farooque, A.A.; Zaman, Q.U.; Esau, T.J.; Chang, Y.K.; Schumann, A.W.; Jameel, W. Influence of wild blueberry fruit yield, plant height, and ground slope on picking performance of a mechanical harvester: Basis for automation. *Appl. Eng. Agric.* **2017**, *33*, 655–666. [CrossRef]
42. Wang, L.; Zhou, X.; Zhu, X.; Dong, Z.; Gou, W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop Sci. Soc. China Inst. Crop Sci. CAAS* **2016**, *4*, 212–219. [CrossRef]
43. Gandhi, N.; Armstrong, L.J.; Petkat, O.; Tripathy, A.K. Rice crop yield prediction in India using support vector machines. In Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016; pp. 1–5.
44. Palanivel, K.; Surianarayanan, C. An approach for prediction of crop yield using machine learning and big data techniques. *Int. J. Eng. Technol.* **2019**, *10*, 110–118. [CrossRef]
45. Obsie, E.Y.; Qu, H.; Drummond, F. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Comput. Electron. Agric.* **2020**, *178*, 105778. [CrossRef]
46. Dang, C.; Liu, Y.; Yue, H.; Qian, J.; Zhu, R. Autumn Crop Yield Prediction using Data-Driven Approaches:—Support Vector Machines, Random Forest, and Deep Neural Network Methods. *Can. J. Remote Sens.* **2021**, *47*, 162–181. [CrossRef]
47. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996; Volume 9.
48. Pan, Y.; Jiang, J.; Wang, R.; Cao, H. Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 169–178. [CrossRef]
49. Üstün, B.; Melssen, W.J.; Buydens, L.M. Facilitating the application of Support Vector Regression by using a universal Pearson VII function-based kernel. *Chemom. Intell. Lab. Syst.* **2006**, *81*, 29–40. [CrossRef]