safety

MDPI

Article

Using Machine Learning to Understand Injuries in Female Agricultural Operators in the Central United States

Cheryl L. Beseler * and Risto H. Rautiainen

Department of Environmental, Agricultural and Occupational Health, University of Nebraska Medical Center, Omaha, NE 68198-4388, USA; rrautiainen@unmc.edu
* Correspondence: chbeseler@unmc.edu

Abstract: The number of women choosing agriculture as an occupation is increasing. Agriculture is dangerous work, and women are at risk of serious injury, but the research on injuries in females is sparse. Women perform different types of farmwork and have different exposures than men. Studies have not assessed injury in a large group of female agricultural operators. In this study, we used XGBoost, a machine learning algorithm, and logistic regression to examine 17 factors hypothesized to be associated with injury in 1529 farm and ranch women. The sample was split into a training group of 1070, and the results were replicated in a test group of 459. The model accuracy was 88%. We compared the results of XGBoost to those of the logistic regression models and computed odds ratios to estimate effect sizes. We found that the two methods generally agreed. XGBoost identified the total number of musculoskeletal symptoms, age, sleep deprivation, high work-related stress, and exposure to respiratory irritants as being important to injury. The multivariate logistic regression model identified higher income, higher stress, younger age, and number of musculoskeletal symptoms as being significantly associated with injury. The analysis highlights the importance of musculoskeletal disorders and work strain to injury in women.

Keywords: agriculture; women; agricultural injuries; XGBoost; injury risk factors; machine learning

Academic Editor: Raphael Grzebieta

Received: 10 December 2024
Revised: 10 January 2025
Accepted: 14 January 2025
Published: 20 January 2025

Citation: Beseler, C.L.; Rautiainen, R.H. Using Machine Learning to Understand Injuries in Female Agricultural Operators in the Central United States. *Safety* **2025**, *11*, 9. https://doi.org/10.3390/safety11010009

1. Introduction

An increasing number of women have become primary agricultural operators on farms and ranches in the United States over the past several decades. The number of farms with female producers increased by 23% between 2012 and 2017, and the number of farms and ranches whose primary operator was female grew by 27% [1]. As of 2017, there were 1.2 million female operators, which is 36% of the 3.4 million agricultural operators in the U.S. [1]. Between 2017 and 2022, the number of female producers in the Midwest grew by 4%, and the trend is likely to continue [2]. However, it is necessary to consider the changes the USDA made to the 2017 Census of Agriculture that extends the number of reported farm operators from three to four but does not distinguish between principal operator and other operator on the farm [3]. This change may add to the number of reported women working in agriculture.

A 2004 study in Kansas identified three reasons for the increased number of women in agriculture after 1997: (1) increased demand for organic and local products; (2) growth in the number of smaller farms; and (3) a greater acceptance of women as principal operators [4]. Although there may be a weakening of the stereotypes for female farmers, women

are moving into what remains a male-dominated occupation where they often feel like outsiders [5,6].

The increase in the number of women involved in food production is changing the agricultural landscape, and economists have been working to understand how. Female operators approach farming differently than males. They show a greater interest in sustainable farming [7], sell products more locally [8], are more likely to be involved in agritourism [9], and earn 40% less in farm income than their male counterparts [10]. A recent study using a Bayesian spatial analysis at the county level showed that a greater number of female farm operators in a county translated to more small business creation, improved life expectancy, and a reduction in the poverty rate [11].

As more women move into agriculture as an occupation, they, too, will be at an increased risk of work-related injuries. Injuries in female agricultural operators are understudied. Most reports focus on male operators because they are the majority of those working in agriculture [12]. A 2004 report found that gender is an important factor in both fatal and non-fatal agricultural injury [13]. Men were eleven times more likely to experience a fatality compared to women and were more likely to be injured by machinery or struck by an object, whereas women were more likely to be injured by an animal. However, a report using five years of Finish insurance data comparing males and females in agricultural occupations found that different work exposures and characteristics partly explain differences in reported injury rates [14]. Comparing male injury rates to female injury rates might be uninformative, and it may be best to study female operators without comparing them to their male counterparts.

To our knowledge, no studies have examined farm characteristics and work exposures in a large sample of agricultural female operators. Injury and risk factor data are nearly always collected using a cross-sectional methodology and are used to make decisions about safety and health priorities. Traditional statistical methods and newer machine learning methods each have their strengths and weaknesses. Using both methods together might provide greater information when exploring the data used to set prevention priorities.

This report examined the factors associated with injury in females working on a farm or ranch. We used two approaches. First, we used Extreme Gradient Boosting (XGBoost), a recent machine learning decision tree approach which has been shown to outperform other machine learning methods, including Random Forests and neural networks [15–17]. XGBoost makes no assumptions about the data such as the independence of the observations, handles missing values, and is unaffected by high correlations among variables while capturing non-linear relationships [18]. We have previously seen that several of our variables show non-linearity across age [19]. Machine learning methods are exploratory; the solution suggested by the method cannot be shown by mathematics to be optimal, but it can help identify patterns in the data and motivate hypotheses using the ability to reliably make predictions using new data. The goal of these algorithms is to calculate the error in the model as each variable is selected from the other variables. Variables that improve the accuracy of the prediction are considered the most important. Second, we used logistic regression to analyze the same set of data, compare the results to XGBoost, and estimate effect sizes. A drawback of methods such as XGBoost is that they do not provide effect estimates or p-values. Recent studies have shown that machine learning methods such as XGBoost outperform logistic regression in predicting the mortality of traumatic brain injury (Wang et al., 2022 [20]), injuries in hockey players (Luu et al., 2020 [21]), cardiovascular disease risk (Xi et al., 2022 [22]), non-specific neck pain (Liew et al., 2022 [23]), and next-season baseball player injuries (Karnuta et al., 2020 [24]), although a meta-analysis of studies predicting acute kidney injury found that they performed similarly (Song et al., 2021 [25]). A systematic review of 71 studies showed no performance improvement by

machine learning methods compared to logistic regression for clinical prediction [26]). It likely depends on the number of variables in the model, the linearity of the relationships between them, and other mathematical relationships in the data that determine whether machine learning models outperform logistic regression models.

In previous analyses of both men and women, important factors in experiencing a work-related agricultural injury included high work-related stress and musculoskeletal symptoms [19]. In a report using a classification tree, injury was an important classifier in musculoskeletal symptoms, sleep deprivation, high work-related stress, and exhaustion [27]. The causal relationships between these characteristics of agricultural work cannot be established in cross-sectional data, but they are likely to be complex. In this study, we examine only women agricultural operators and use an algorithm published in 2016 that outperforms other machine learning methods to compare the features that distinguish women who have reported a work-related injury from those who have not [18]. We expect that the same factors we have seen before will again be important. High work-related stress, exhaustion, sleep deprivation, and musculoskeletal symptoms will be primary contributors to injury in agricultural women. We expected to see differences after comparing the results of XGBoost and logistic regression due to differences in the algorithms.

2. Materials and Methods

2.1. Sample

The sample derives from three cross-sectional surveys administered in the spring and summer of 2018, 2020, and 2023 to farm and ranch operators in a seven-state region (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota). Each was an unduplicated sample such that no operators were included in more than one survey. The paper-based survey was distributed to a random sample of approximately 2500 farm and ranch operations in each of the seven states in each of the three years. The Farm and Ranch Health and Safety Survey (FRHSS), consisting of 29 questions, was created by the Central States Center for Agricultural Safety and Health (CS-CASH). The survey focused on injuries, chronic health outcomes, work-related exposures, and the use of personal protective equipment for up to three main operators working on a farm or ranch. All responses were received by mail and the data were entered into the Research Electronic Data Capture (REDCap 14.9.6) software. The study was determined to be exempt from human subject research by the University of Nebraska Medical Center Institutional Review Board (No. 452-11-EX).

The sampling frame was provided by Farm Market iD (FMiD) (2018 and 2020 surveys) and US Farm Data (2023 survey), each using a stratified random sample. FMiD and US Farm Data are for-profit organizations that create farm databases from the United States Department of Agriculture (USDA) annual surveys, remote sensing data, and other public and private sources. The FMiD/US Farm Data databases cover approximately 95% of agricultural operations in the United States. Farm and ranch operators with an email address and an estimated gross farm income of at least USD 5000 were eligible to participate. Operators were those who ran a farm or ranch business and were not considered farm workers. The 2018 random sample was sent out in May 2018 with one follow-up request in June and resulted in a response rate of 19%, including 3268 farms and ranches and 4423 individual operators. Similarly, the 2020 survey was sent out in March and the follow-up request was mailed out in June. The 2736 farms that responded represented 3492 individual farmers and ranchers (response rate of 16%). In the 2023 survey, 1770 farms responded out of 17,497, representing 2367 operators, for a response rate of 10.1%. We included only females who reported being a primary, secondary, or tertiary operator on a farm or ranch and were 18 years of age or older.

2.2. Measures

We included 17 measures that were hypothesized to be associated with experiencing an injury. We included age because it has been shown to be a factor in injury risk [19,28,29]. Work characteristics included whether a woman worked on a farm, ranch, or both; whether her primary occupation was an agricultural operator or other occupation; and whether she was a principal, secondary, or tertiary operator on the operation. The percentage of time spent working on the farm or ranch was a categorical variable (100, 75–99, 50–74, 25–49, 0–24). These four variables assessed the work responsibilities on the farm and the time available to accomplish the tasks. We included two diagnosed medical conditions that might inhibit work productivity, namely respiratory and skin disorders (yes or no). We summed up nine possible regions of the body (neck, shoulder, upper back, elbows, wrists/hands, low back, hips/thighs, knees, ankle/feet) that might experience musculoskeletal discomfort based on the hypothesis that the more discomfort in the body, the greater the impairment an operator might experience, and impairment could increase the risk of injury. We also added binary variables (yes/no) for respiratory exposures, exposure to noise, chemical exposures to skin, and musculoskeletal exposures (working in awkward positions causing strain). We included a protective factor asking about engaging in prevention exercises to prevent musculoskeletal disorders (yes/no). Lastly, we included measures of work strain, including work-related sleep deprivation, high work-related stress, and exhaustion, all of which were binary variables.

The outcome variable, the supervisor in machine learning language, was asked using the question "How many farm-related injuries occurred to each operator during the past 12 months?", with potential responses of none, one, two, or three or more. The responses were categorized into a binary variable, where 0 was coded as not having any injury and 1 was coded when respondents reported at least one injury.

2.3. Statistical Analysis

2.3.1. Sample

Descriptive summary measures were calculated for the entire sample and for the training data and test data separately. Chi-square tests were used to assure that there were no statistically significant differences after randomly dividing the sample data into training and test datasets.

2.3.2. XGBoost

To conduct the machine learning analysis, the three sets of survey data were combined into a single data file, and the data were randomly shuffled to introduce randomness into the order of the observations and randomly split into 70% to be used for training and 30% for testing. XGBoost was used to estimate a model with two boosting iterations, regularization, and three early stopping rounds to reduce overfitting, and a logistic objective function with injury (0 = no, 1 = yes) as the supervisor. We applied the specified model to the training data and then used the trained model with the new test data to estimate the misclassification rate. The accuracy rate was calculated by comparing how often the actual outcome in the testing data matched the outcome value predicted by the model with the new test data. The XGBoost model was subsequently tuned by setting the maximum depth of each decision tree to three with ten boosting rounds. Variable importance was calculated using the improvement in model accuracy with each selected variable.

2.3.3. Logistic Regression

In the next step, we tested all 17 variables in univariable and multivariable logistic regression models to assess their association with injury. We calculated the variance inflation factor to check for multicollinearity and estimated a final set of variables that

explained injury as measured by the odds ratios and 95% confidence intervals. To compare the logistic regression results to the XGBoost results, the variables which were shown to be important in the tree-based model were used in the univariable and multivariable logistic regression models, and effect sizes were obtained. The results of the two models were compared based on the variables selected and their odds ratios and 95% confidence intervals. Comparisons were also made between the original logistic regression model with all 17 variables and the XGBoost results.

3. Results

3.1. Sample

The mean age of the 1529 women was 59.7 (SD = 12.2). Most of them were secondary operators (79.7%), indicating they were the spouse of the primary operator. Further, 56.1% reported that farming/ranching was their primary occupation and that they spent at least half of their work time on the operation. The majority reported working primarily on a farming operation (61%), and about a quarter were both farmers and ranchers. The gross farm incomes of their operations covered a wide range, with 16% being under USD 50,000 per year and 30% earning at least USD 500,000. Table 1 shows the detailed characteristics of the sample.

Table 1. Demographics and farm characteristics of sample used in XGBoost analysis.

Variable	Total Sample n = 1529 n (%)	Training Sample n = 1070 n (%)	Test Sample n = 459 n (%)
Injury (outcome)			
Yes	156 (10.2)	107 (10.0)	49 (10.7)
No	1373 (89.8)	963 (90.0)	410 (89.3)
Operator			
Primary	202 (13.2)	147 (13.7)	55 (12.0)
Secondary	1218 (79.7)	849 (79.4)	369 (80.4)
Tertiary	109 (7.13)	74 (6.92)	35 (7.63)
Primary occupation			
Farm/ranch work	847 (56.1)	589 (55.8)	258 (57.0)
Other	662 (43.9)	467 (44.2)	195 (43.0)
Percentage time working on operation			
100%	385 (25.9)	269 (25.9)	116 (26.1)
75–99%	227 (15.3)	158 (15.2)	69 (15.5)
50–74%	223 (15.0)	150 (14.4)	73 (16.4)
25–49%	332 (22.4)	234 (22.5)	98 (22.0)
0–24%	318 (21.4)	229 (22.0)	89 (20.0)
Farm or ranch			
Farm	926 (60.6)	654 (61.1)	272 (59.3)
Ranch	231 (15.1)	158 (14.8)	73 (15.9)
Both	372 (24.3)	258 (24.1)	114 (24.8)
Estimated revenue			
<50,000	245 (16.1)	179 (16.8)	66 (14.5)
50,000–100,000	143 (9.41)	105 (9.87)	38 (8.33)
100,000–200,000	233 (15.3)	155 (14.6)	78 (17.1)
200,000–300,000	172 (11.3)	127 (11.9)	45 (9.87)
300,000–400,000	150 (9.87)	99 (9.30)	51 (11.2)
400,000–500,000	118 (7.76)	82 (7.71)	36 (7.89)
500,000–1,000,000	287 (18.9)	193 (18.1)	94 (20.6)
1,000,000–2,000,000	128 (8.42)	92 (8.65)	36 (7.89)
2,000,000–3,000,000	27 (1.78)	19 (1.79)	8 (1.75)
3,000,000–5,000,000	12 (0.79)	10 (0.94)	2 (0.44)
>5,000,000	5 (0.33)	3 (0.28)	2 (0.44)

Table 1. *Cont.*

Variable	Total Sample n = 1529 n (%)	Training Sample n = 1070 n (%)	Test Sample n = 459 n (%)
Respiratory condition			
Yes	417 (27.3)	282 (26.4)	135 (29.4)
No	1112 (72.7)	788 (73.6)	324 (70.6)
Skin disorder			
Yes	294 (19.2)	217 (20.3)	77 (16.8)
No	1235 (80.8)	853 (79.7)	382 (83.2)
High work-related stress			
Yes	338 (22.1)	226 (21.1)	112 (24.4)
No	1191 (77.9)	844 (78.9)	347 (75.6)
Sleep deprivation			
Yes	293 (19.2)	192 (17.9)	101 (22.0)
No	1236 (80.8)	878 (82.1)	358 (78.0)
Exhaustion			
Yes	345 (22.6)	230 (21.5)	115 (25.1)
No	1184 (77.4)	840 (78.5)	344 (74.9)
Musculoskeletal discomfort exposure			
Yes	1070 (70.0)	747 (69.8)	323 (70.4)
No	459 (30.0)	323 (30.2)	136 (29.6)
Noise exposure *			
Yes	990 (64.7)	675 (63.1)	315 (68.6)
No	539 (35.3)	395 (36.9)	144 (31.4)
Respiratory exposures			
Yes	786 (51.4)	548 (51.2)	238 (51.8)
No	743 (48.6)	522 (48.8)	221 (48.2)
Skin exposures			
Yes	1042 (68.2)	717 (67.0)	325 (70.8)
No	487 (31.8)	353 (33.0)	134 (29.2)
Use MSD prevention techniques			
Yes	1218 (79.7)	851 (79.5)	367 (80.0)
No	311 (20.3)	219 (20.5)	92 (20.0)
Continuous variables	Mean (SD)	Mean (SD)	Mean (SD)
Age	59.7 (12.2)	60.1 (12.2)	58.9 (12.0)
Number of musculoskeletal disorders	1.26 (1.61)	1.25 (1.62)	1.28 (1.59)

* p-value = 0.04 for differences in test and training samples.

About 20% of the women who responded reported high stress, sleep deprivation, and exhaustion (Table 1). The majority reported being exposed to loud noise, respiratory and skin irritants, and work postures that could result in musculoskeletal symptoms (MSSs). Nearly 80% reported taking preventive measures to reduce their risk of musculoskeletal discomfort, although the mean number of body areas where they experienced discomfort was greater than one. Approximately 10% (n = 156) of the women reported having had an injury. Of those who reported an injury, 105 (6.81%) sought care from a doctor and 16 (1.04%) sought care at a hospital. Although 65 (38%) did not lose time from work, 24 (14%) lost at least 30 days from work.

3.2. XGBoost

The total sample of 1529 female agricultural operators was split into 1070 in the training data and 459 in the testing data. With 17 variables, it would not be surprising to see a spurious statistical association in these groups, and this occurred with noise exposure (X^2 = 4.32, p = 0.04). None of the other comparisons between the training and testing groups were close to statistically significant. The XGBoost results showed a misclassification rate of 12.85% before

tuning the model and 11.55% after tuning the model. This represents an accuracy rate of better than 88% after tuning the initial model. The most important contributors to explaining injury as measured by the improvement in model accuracy were the total number of body areas affected by MSSs (0.60), age (0.29), sleep deprivation (0.05), high work-related stress (0.04), and exposure to respiratory irritants (0.02). None of the other 12 variables contributed to improving the classification of those with and without an injury.

3.3. Logistic Regression

Fifteen of the seventeen variables were significantly associated with injury in the univariable logistic regression models (Table 2). The only exceptions were the role of the operator (primary, secondary, tertiary) and whether the operation was a farm, ranch, or both ($p = 0.05$). The criteria for selecting variables for the XGBoost model were shown to be valid choices based on the univariable logistic regression models. In the multivariable model, which included all 17 variables, higher incomes were associated with more reported injuries, as were high work-related stress, younger age, and the number of MSSs. The variance inflation factor for each variable was less than 2.5, so collinearity should not have been a problem in the model. Only four of the seventeen variables remained statistically significant in the final adjusted logistic regression model (Table 2).

Table 2. Odds ratios (95% confidence intervals) from logistic regression models of 17 variables hypothesized to be associated with injury in 1529 female agricultural operators.

Variable	Univariable OR (95% CI)	Multivariable OR (95% CI)
Operator		
Primary	Reference	Reference
Secondary	0.91 (0.56, 1.47)	0.78 (0.45, 1.34)
Tertiary	1.01 (0.48, 2.13)	0.72 (0.31, 1.69)
Primary occupation		
Farm/ranch work	Reference	Reference
Other	0.55 (0.38, 0.78)	0.68 (0.38, 1.22)
Percentage time working on operation		
100%	Reference	Reference
75–99%	1.05 (0.64, 1.73)	1.09 (0.64, 1.85)
50–74%	1.03 (0.63, 1.70)	1.39 (0.78, 2.45)
25–49%	0.82 (0.51, 1.31)	1.50 (0.75, 3.00)
0–24%	0.33 (0.18, 0.61)	0.93 (0.40, 2.17)
Farm or ranch		
Both	Reference	Reference
Farm	0.70 (0.48, 1.02)	0.70 (0.47, 1.06)
Ranch	0.94 (0.57, 1.56)	0.86 (0.49, 1.49)
Estimated revenue		
1.11 (1.03, 1.18)		
No	Reference	Reference
Yes	1.78 (1.26, 2.51)	1.24 (0.84, 1.84)
Diagnosed skin disorder		
No	Reference	Reference
Yes	1.70 (1.17, 2.48)	1.45 (0.94, 2.22)
High work-related stress		
No	Reference	Reference
Yes	2.97 (2.10, 4.19)	1.61 (1.03, 2.52)
Sleep deprivation		
No	Reference	Reference
Yes	2.43 (1.70, 3.48)	1.06 (0.67, 1.68)
Exhaustion		
No	Reference	Reference
Yes	2.54 (1.80, 3.59)	0.96 (0.60, 1.52)

Table 2. *Cont.*

Variable	Univariable OR (95% CI)	Multivariable OR (95% CI)
Work positions leading to musculoskeletal discomfort		
No	Reference	Reference
Yes	3.40 (2.08, 5.57)	1.80 (0.94, 3.42)
Noise exposure		
No	Reference	Reference
Yes	1.85 (1.26, 2.72)	1.03 (0.64, 1.66)
Respiratory exposures		
No	Reference	Reference
Yes	2.23 (1.57, 3.18)	0.97 (0.62, 1.52)
Skin exposures		
No	Reference	Reference
Yes	2.19 (1.44, 3.33)	0.94 (0.56, 1.57)
Use MSD prevention techniques		
No	Reference	Reference
Yes	2.80 (1.59, 4.92)	1.84 (0.93, 3.62)
Age	0.97 (0.96, 0.99)	0.98 (0.96, 0.99)
Number of musculoskeletal symptoms	1.39 (1.27, 1.51)	1.25 (1.13, 1.40)

3.4. Comparing XGBoost to Logistic Regression

Table 3 shows the four variables deemed important from the XGBoost model when used in the univariable and multivariable logistic regression models. All variables were statistically significant in the univariable analysis. Although the total number of MSSs was of greatest importance in the XGBoost model, it was of secondary importance in the logistic regression model. However, the 95% CI shows more certainty around the musculoskeletal variable than it does around the stress variable. Respiratory exposure was not statistically significant in the multivariable model, and it was the least important in the XGBoost model. However, it had a strong association with injury in the univariable model, and it is not clear which other variable in the model would have impacted this OR in a downward direction.

Table 3. Odds ratios and 95% confidence intervals (CIs) from logistic regression of variables identified as important in XGBoost machine learning algorithm.

Variable	Univariable OR (95% CI)	Multivariable OR (95% CI)
Musculoskeletal symptoms	1.39 (1.27, 1.51)	1.28 (1.17, 1.41)
Age	0.97 (0.96, 0.99)	0.98 (0.96, 0.99)
Sleep deprivation	2.43 (1.70, 3.48)	1.14 (0.74, 1.75)
High work-related stress	2.97 (2.10, 4.19)	1.57 (1.03, 2.39)
Respiratory exposures	2.23 (1.57, 3.18)	1.42 (0.97, 2.09)

XGBoost did not find income to be an important predictor, and the multivariable logistic regression model did not find sleep deprivation or respiratory exposures to be associated with injury. The models agree that stress, MSSs, and age separate injured women from uninjured women.

4. Discussion

It is reassuring that traditional logistic regression models do not differ greatly from machine learning methodologies. Five of seventeen variables were identified by the XGBoost algorithm, and four of seventeen were significant in the logistic regression model. XGBoost and logistic regression agreed that high work-related stress and MSSs are related to experi-

encing an agricultural injury in women, but there were differences in the results. XGBoost identified the total number of MSSs in various body regions to be the most important factor associated with injury. Since the method does not infer a causal direction, it might be that MSSs are a result of an injury and therefore the model selected them as being the most important determinant of injury. It might also be that MSSs create vulnerability to injury. The relationship is possibly reciprocal. More work should be conducted to disentangle this association.

The second most important feature of injury in the XGBoost model was younger age, although this variable was not as important based on the odds ratio from the logistic regression. Sleep deprivation, high work-related stress, and respiratory exposures were also important in XGBoost, but the logistic regression retained only stress in the multivariable model. The relationship between stress and sleep may be complicated and not captured in traditional models of injury. This complexity may be best captured by machine learning algorithms rather than maximum likelihood estimation.

Although both XGBoost and logistic regression are classification methods where the outcome is a binary variable and the features are a set of explanatory variables, the algorithms are completely different. Whereas maximum likelihood adjusts all variables simultaneously to find the value of the parameters that maximize the function (regression equation), gradient descent methods find a minimum function that reduces the error at each iteration so it learns from the previous iterations. Gradient descent is used to iteratively search a grid that will minimize the loss function. The result is that logistic regression models are more likely to be influenced by other variables in the model and methods such as XGBoost less so. This can create differing results, but both can be informative.

On any given agricultural operation, the possible set of factors will influence each other in ways that are unique to a specific operation. This could make gradient descent methods more useful because they are likely to look at each variable independent of every other variable while iteratively correcting for classification errors that occur. The entire approach is designed to increase the accuracy of prediction, which is important when trying to understand how injuries happen. Knowledge of agricultural safety and health must be employed to interpret the results when they may not appear causal but are reproducible.

Logistic regression models do not handle irregular patterns in the data, including non-linearity, inconsistent interactions, and collinearity between variables. XGBoost does a better job than other tree-based methods or even neural networks because of its ensemble approach to building trees, removing uninformative features, and creating linear combinations that better represent non-continuous data [16]. Tree-based methods adapt well to tabular data with binary or ordinal variables.

The results highlight the important need for more work on musculoskeletal disorders and their relationship to stress and sleep deprivation in terms of the risk of injury. The frequency of women reporting any MSS was high in this sample (62.9%). The reported frequency of women using preventive techniques to prevent musculoskeletal symptoms showed that 32.9% of those who did not report any discomfort did not use any prevention techniques; however, 56.2% of those with no discomfort did report using preventive techniques. The 24% difference in these numbers suggests that, possibly, using prevention techniques reduces musculoskeletal symptoms, or that the 32.2% will start performing exercises to prevent symptoms should they begin developing symptoms.

Other interesting results that might be unique to women operators include the lack of association between injury and their role in the operation; whether the operation was a farm, ranch, or both; whether they worked off the farm or ranch; and time spent working on the operation. A study of large-machinery-related injuries in 7420 male and female agricultural operators in the Midwest found that a higher risk of injury was associated with being male,

being older in age, having a history of a prior injury, and working an increasing number of hours on the operation [30]. An older case–control study of strictly female operators found only two significant risk factors for injury, namely working an increasing number of hours and the presence of bulls on the operation [31]. Risk factors have possibly changed for women working on farms and ranches since these studies were published, but a woman's work responsibilities and time spent working on the operation did not appear to be important risk factors for injury. It is an interesting question whether a history of a prior injury creates vulnerability to a subsequent injury due to having musculoskeletal discomfort.

The limitations of this study are several, including the limited number of risk factors asked about, the self-selected nature of the survey data, and the potential biases present when asking people to recall events over the past year. If an injury is serious, farmers are likely to remember it. It is also possible that those with an injury are more likely to be interested in responding to a survey about injuries. The number of important factors associated with experiencing an injury was low in this study, and many others were not examined due to not being asked about in our survey or being unknown. The cross-sectional surveillance data do not allow us to draw causal inferences.

5. Conclusions

We found minor differences between the results from the machine learning method used in XGBoost and traditional logistic regression. XGBoost identified a few additional variables that logistic regression did not, but they were not highly important in the machine learning model. The models agreed that work-related stress and musculoskeletal symptoms were important modifiable variables that either are due to injury or might cause someone to be more vulnerable to injury. Although it may be a difficult task to intervene to reduce the stress that agricultural women are feeling, sleep is a modifiable target that would likely impact both perceived stress and a person's reaction to stress. Although we did not identify any novel aspect of farmwork where we can intervene to reduce injury, we have more work to carry out to understand how to reduce musculoskeletal pain and discomfort and how this might reduce the stressors female operators are experiencing.

Author Contributions: Conceptualization, C.L.B.; Data curation, R.H.R.; Formal analysis, C.L.B.; Funding acquisition, R.H.R.; Methodology, C.L.B.; Writing—original draft, C.L.B.; Writing—review and editing, R.H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the CDC/NIOSH under grant number U54 OH010162 for the Central States Center for Agricultural Safety and Health.

Institutional Review Board Statement: This study was determined to be exempt from human subject research by the University of Nebraska Medical Center Institutional Review Board (No. 452-11-EX). Ethical review and approval were waived for this study due to this being surveillance and not human subject research involving humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be obtained by contacting the first author.

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

References

1. USDA. National Agricultural Statistics Service. Female Producers. 2019. Available online: https://www.nass.usda.gov/Publications/Highlights/2019/2017Census_Female_Producers.pdf (accessed on 10 November 2024).
2. USDA. National Agricultural Statistics Service. 2022 Census of Agriculture. Available online: https://www.nass.usda.gov/AgCensus/ (accessed on 10 November 2024).

3. Pilgeram, R.; Dentzman, K.; Lewin, P.; Conley, K. How the USDA changed the way women farmers are counted in the Census of Agriculture. *Choices* **2020**, *35*, 1–10.
4. Ball, J.A. She works hard for the money: Women in Kansas agriculture. *Agric. Hum. Values* **2014**, *31*, 595–605. [CrossRef]
5. Kubik, W.; Moore, R.J. Health and well-being of farm women: Contradictory roles in the contemporary economy. *J. Agric. Saf. Health* **2005**, *11*, 249–256. [CrossRef] [PubMed]
6. Pilgeram, R.; Dentzman, K.; Lewin, P. Women, race and place in US Agriculture. *Agric. Hum. Values* **2022**, *39*, 1341–1355. [CrossRef] [PubMed]
7. Trauger, A. "Because they can do the work": Women farmers in sustainable agriculture in Pennsylvania, USA. *Gend. Place Cult.* **2004**, *11*, 289–307. [CrossRef]
8. Trauger, A.; Sachs, C.; Barbercheck, M.; Brasier, K.; Kiernan, N.E. "Our market is our community": Women farmers and civic agriculture in Pennsylvania, USA. *Agric. Hum. Values* **2010**, *27*, 43–55. [CrossRef]
9. Schmidt, C.; Goetz, S.J.; Tian, Z. Female farmers in the United States: Research needs and policy questions. *Food Policy* **2021**, *101*, 102039. [CrossRef]
10. Fremstad, A.; Paul, M. Opening the farm gage to women? The gender gap in US agriculture. *J. Econ. Issues* **2020**, *54*, 124–141. [CrossRef]
11. Schmidt, C.; Deller, S.C.; Goetz, S.J. Women farmers and community well-being under modeling uncertainty. *Appl. Econ. Perspect. Policy* **2024**, *46*, 275–299. [CrossRef]
12. Prater, L.F. Health Needs of Women in Ag Overlooked. *Successful Farming*, 18 May 2022. Available online: https://www.agriculture.com/family/health-safety/health-needs-of-women-in-ag-overlooked (accessed on 24 November 2024).
13. Dimich-Ward, H.; Guernsey, J.R.; Pickett, W.; Renie, D.; Hartling, L.; Brison, R.J. Gender differences in the occurrence of farm related injuries. *Occup. Environ. Med.* **2004**, *61*, 52–56.
14. Karttunen, J.P.; Rautiainen, R.H.; Quendler, E. Gender division of farm work and occupational injuries. *J. Agric. Saf. Health* **2019**, *25*, 117–127. [CrossRef] [PubMed]
15. Kossen, J.; Band, N.; Lyle, C.; Gomez, A.N.; Rainforth, T.; Gal, Y. Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *arXiv* **2021**, arXiv:2106.02584.
16. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based methods still outperform deep learning on tabular data? *arXiv* **2022**, arXiv:2207.08815v1.
17. Montomoli, J.; Romeo, L.; Moccia, S.; Bernardini, M.; Migiorelli, L.; Berardin, D.; Donati, A.; Carsetti, A.; Bocci, G.; Garcia, P.D.W.; et al. RISC-19-ICU Investigators. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA Score at ICU admission in COVID-19 patients. *J. Intensive Med.* **2021**, *1*, 110–116. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
19. Beseler, C.L.; Rautiainen, R.H. Injury, musculoskeletal symptoms, and stress as a function of aging in agricultural operators. *Workplace Health Saf.* **2023**, *17*, 597–605. [CrossRef]
20. Wang, R.; Wang, L.; Zhang, J.; He, M.; Xu, J. XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. *World Neurosurg.* **2022**, *163*, e617–e622. [CrossRef]
21. Luu, B.C.; Wright, A.L.; Haeberle, H.S.; Karnuta, J.M.; Schickendantz, M.S.; Makhni, E.C.; Nwachukwu, B.U.; Williams, R.J.; Ramkumar, P.N. Machine learning outperforms logistic regression analysis to predict next-season NHL player injury. An analysis of 2322 players from 2007 to 2017. *Orthop. J. Sports Med.* **2020**, *8*, 2325967120953404. [CrossRef]
22. Xi, Y.; Wang, H.Y.; Sun, N.L. Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. *Front. Cardiovasc. Med.* **2022**, *9*, 1025705. [CrossRef]
23. Liew, B.X.W.; Kovacs, F.M.; Rügamer, D.; Royuela, A. Machine learning versus logistic regression for prognostic modelling in individuals with non-specific neck pain. *Eur. Spine J.* **2022**, *31*, 2082–2091. [CrossRef]
24. Karnuta, J.M.; Luu, B.C.; Haeberle, H.S.; Saluan, P.M.; Frangiamore, S.J.; Stearns, K.L.; Farrow, L.D.; Nwachukwu, B.U.; Verma, N.N.; Makhni, E.C.; et al. Machine learning outperforms regression analysis to predict next-season major league baseball player injuries. *Orthop. J. Sports Med.* **2020**, *8*, 2325967120963046. [CrossRef]
25. Song, X.; Liu, X.; Liu, F.; Wang, C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *Int. J. Med. Inform.* **2021**, *151*, 104484. [CrossRef] [PubMed]
26. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Calster, B.V. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [CrossRef] [PubMed]
27. Chengane, S.; Beseler, C.L.; Duysen, E.G.; Rautiainen, R.H. Occupational stress among farm and ranch operators from a seven state surveillance system in the midwestern United States. *BMC Public Health* **2021**, *21*, 2076. [CrossRef] [PubMed]

28. Gorucu, S.; Murphy, D.J.; Kassab, C. A multi-year analysis of fatal farm and agricultural injuries in Pennsylvania. *J. Agric. Saf. Health* **2015**, *21*, 281–298.
29. Jadhav, R.; Lander, L.; Achutan, C.; Haynatzki, G.; Rajaram, S.; Patel, K.; Rautiainen, R. Review and meta-analysis of emerging risk factors for agricultural injury. *J. Agromed.* **2016**, *21*, 1–14. [CrossRef]
30. Reiner, A.M.; Gerberich, S.G.; Ryan, A.D.; Mandel, J. Large machinery-related agricultural injuries across a five-state region in the Midwest. *J. Occup. Environ. Med.* **2016**, *58*, 154–161. [CrossRef]
31. Stueland, D.T.; Lee, B.C.; Nordstrum, D.L.; Layde, P.M.; Wittman, L.M.; Gunderson, P.D. Case-control study of agricultural injuries to women in Central Wisconsin. *Women Health* **1997**, *25*, 91–103. [CrossRef]