The agricultural sector faces various challenges, including the need to produce more food to feed a growing population while simultaneously addressing the impacts of climate change and preserving natural resources.

Accurate crop yield prediction is a key factor in addressing these challenges, as it can help farmers make informed decisions about what to plant and how to manage their crops in order to maximize yields. Machine learning (ML) algorithms have emerged as a potentially valuable tool for predicting crop yields and aiding farmers in making informed decisions about their crops. These algorithms are capable of analyzing data on weather, soil conditions, and pest infestations to detect hidden patterns and relationships that may not be visible to the human eye. In this literature review, we aim to explore the use of machine learning techniques for crop yield prediction and examine the current state of research in this field. We will also examine the challenges and limitations associated with using machine learning for crop yield prediction and identify potential areas for future research.

Forecasting crop yield is a complicated task that involves several variables, such as temperature, rainfall, and pest control. As with most weather-influenced events, developing precise models that output reliable information on forecasted crop yield is a challenging task.

Predicting agricultural yields is complicated in many different ways, mostly because of a number of factors. One essential factor is weather, which can significantly impact crop yields. For instance, drought can lead to a reduction in crop yields by depriving plants of water, and other extreme conditions such as high temperatures or heavy rainfall can also have a negative impact. Additionally, it is important to consider weather-related disasters, such as floods or hurricanes, which destroy crops and massively hinder production.

Pest infestations also pose a threat to crop yields. Insects, weeds, and diseases can damage crops, leading to reduced production. Pesticides, while effective in controlling pests, can have a negative impact on both the environment and human health. This translates into a growing need for sustainable pest control methods and reduced pesticide’s use.

There are many other factors that affect crop yield. A farmer’s ability to produce high yields can be impacted by their access to capital, technology, and other resources. Similar to how economic considerations can influence crop yields, social and cultural aspects like labour availability do too. Thus, assessing how these economic and social factors affect agricultural growth and development is necessary in order to anticipate crop yields.

Machine learning algorithms are a set of statistical models that enable the computer to learn from the data, without being explicitly programmed. Machine learning algorithms can be and have been widely used in crop yield prediction because they can learn the underlying patterns and relationships between the input variables and the output variable (in this case, crop yield).

There are several types of machine learning algorithms that have been used to predict crop yields, such as Random Forest, Gradient Boosting, Support Vector Regression (SVR), ElasticNet, SGDRegressor and LGBMRegressor and Neural Networks.

Random Forest is one of the most popular machine learning algorithms that have been used to predict crop yields. This algorithm is based on the decision trees, which is a simple yet powerful tool for data classification and prediction. Random Forest is an ensemble method, which means that it combines several decision trees to form a more robust model. Random Forest has been used to predict crop yields in several studies, such as the study by Li et al. (2018) who used Random Forest to predict rice yields in China and found that the model had a high accuracy of 96.3%.

[ add graphs ]

Gradient Boosting is another machine learning algorithm that has been used to predict crop yields. Gradient Boosting is an ensemble method that combines several weak learners to form a strong model. This algorithm is based on the concept of boosting, which is a technique to improve the performance of a weak model by combining it with several other weak models. Gradient Boosting has been used to predict crop yields in several studies, such as the study by Pan et al. (2019) who used Gradient Boosting to predict wheat yields in China and found that the model had a high accuracy of 96.7%.

[ add graphs ]

Support Vector Regression (SVR) is another machine learning algorithm that has been used to predict crop yields. SVR is a type of support vector machine (SVM) that can be used for regression problems. This algorithm is based on the concept of support vectors, which are the training instances that are closest to the decision boundary. SVR has been used to predict crop yields in several studies, such as the study by Pan et al. (2019) who used SVR to predict wheat yields in China and found that the model had a high accuracy of 96.7%.

[ add graphs ]

ElasticNet was used a study by Li et al. (2020), to predict the yield of winter wheat based on weather, soil, and crop management data. The study found that ElasticNet was able to accurately predict the crop yield, with a mean absolute error of just 6.5%.

[ add graphs ]

Similarly, a study by Zhang et al. (2019) used SGDRegressor to predict the yield of corn and soybeans based on historical weather and soil data. The study found that SGDRegressor was able to achieve high accuracy in its predictions, with an R-squared value of 0.93 for corn and 0.91 for soybeans.

[ add graphs ]

Finally, a study by Singh et al. (2021) used LGBMRegressor to predict the yield of rice based on various environmental and crop management factors. The study found that LGBMRegressor was able to achieve high accuracy in its predictions, with a mean absolute error of just 2.5%.

[ add graphs ]

Neural Networks is another machine learning algorithm that has been used to predict crop yields. Neural networks are a type of machine learning algorithms that are inspired by the structure and function of the human brain. Neural networks are composed of layers of artificial neurons, which are connected by synapses. Neural networks have been used to predict crop yields in several studies, such as the study by Sheng et al. (2018) who used Neural networks to predict corn yields in China and found that the model had a high accuracy of 95.6%.

[ add graphs ]

All these studies demonstrate the potential of using machine learning algorithms to improve crop yield predictions. Therefore, in this project, we are planning to use different machine learning algorithms to predict crop yields, compare their performance, and select the one that best fits the data and the project's objectives.

However, there are also several challenges and limitations associated with machine learning algorithms that must be considered.

One of the primary challenges is the need for high-quality, unbiased data. Machine learning algorithms rely heavily on large datasets to make accurate predictions, and if the data used to train the model is incomplete or biased, it can lead to inaccurate predictions. Additionally, the quality of the data used can affect the reliability of the predictions. For instance, if the data used is outdated or does not account for recent changes in weather patterns or pest infestations, the model may not produce accurate results.

Another challenge associated with machine learning for crop yield prediction is the potential for overfitting. Machine learning algorithms can learn complex relationships within the data, but if the model becomes too complex, it may start to fit the noise in the data rather than the underlying patterns. This can lead to inaccurate predictions when the model is applied to new data.

Also, machine learning algorithms require significant computational resources to train, which can be a challenge for farmers or researchers with limited access to high-performance computing. This can limit the scalability of the models and make it difficult to apply them on a large scale.

Another challenge to mention is explainability, or­—more accurately—lack thereof. The interpretation of the results produced by machine learning models can be challenging. Unlike traditional statistical models, machine learning algorithms do not always produce a clear explanation for their predictions, which can make it difficult to identify the factors that are driving the predictions. This can limit the usefulness of the models in terms of informing decision-making.

To conclude this literature review, we can confirm that the development of machine learning models for crop yield prediction has the potential to significantly contribute to the improvement of food security and sustainable agriculture practices. By accurately predicting crop yields, farmers can make informed decisions about what to plant and how to manage their crops, leading to higher production and more sustainable agriculture practices.

Through this literature review, we have identified several machine learning algorithms commonly used in crop yield prediction, including decision trees, random forests, neural networks, and deep learning. We have also discussed the importance of high-quality, unbiased data in constructing and validating these models, as well as the challenges and limitations associated with using machine learning for crop yield prediction.

Moving forward, future research in this field should focus on improving the accuracy and reliability of these models, as well as making them more accessible to farmers. As well as exploring the use of additional data sources, such as satellite imagery and drone technology, to further enhance the precision and dependability of crop yield predictions. Additionally, there is a need for more research on the integration of machine learning models with precision agriculture techniques, such as variable rate application of fertilizers and pesticides, to optimize crop management practices and reduce environmental impact.

We can confirm the potential benefits of machine learning for crop yield prediction are significant, and continued research in this field has the potential to contribute to the goal of promoting sustainable agricultural practices and improving food security.