Deep density estimation based on multi-spectral remote sensing data for in-field crop yield forecasting

Liana Baghdasaryan
Intelinair, Inc.
Yerevan, Am
liana@intelinair.com

Razmik Melikbekyan
Intelinair, Inc.
Yerevan, Am
razmik@intelinair.com

Arthur Dolmajain
Intelinair, Inc.
Yerevan, Am
arthur@intelinair.com

Jennifer Hobbs
Intelinair, Inc.
Chicago, IL, USA
jennifer@intelinair.com

Abstract

Yield forecasting has been a central task in computational agriculture because of its impact on agricultural management from the individual farmer to the government level. With advances in remote sensing technology, computational processing power, and machine learning, the ability to forecast yield has improved substantially over the past years. However, most previous work has been done leveraging low-resolution satellite imagery and forecasting yield at the region, county, or occasionally farm-level. In this work, we use high-resolution aerial imagery and output from high-precision harvesters to predict in-field harvest values for corn-raising farms in the US Midwest. By using the harvester information, we are able to cast the problem of yield-forecasting as a density estimation problem and predict a harvest rate, in bushels/acre, at every pixel in the field image. This approach provides the farmer with a detailed view of which areas of the farm may be performing poorly so he can make the appropriate management decisions in addition to providing an improved prediction of total yield. We evaluate both traditional machine learning approaches with hand-crafted features alongside deep learning methods. We demonstrate the superiority of our pixel-level approach based on an encoder-decoder framework which produces a 5.41% MAPE at the field-level.

1. Introduction

Although initially a slow adopter of machine learning and computer vision, agriculture has become an important domain for these approaches. Computer vision is now a key element of agricultural systems to determine crop type [47], count plants [16, 28], guide harvesting robots [22], identify issues like crop stress and weeds [7, 10, 35, 39], and forecast yield [3, 23, 52]. Adoption and extension of these approaches is critical due to the challenges facing global agriculture: the world's population is predicted to reach 9.7billion by 2050 [34], water supply is expected to fall 40% short of global needs by 2030 [32], and climate change produces significant challenges and uncertainty [11].

Crop yield forecasting is a central task in precision agriculture because of its impact on food security, economics, and scientific development. Numerous stakeholders are impacted: farmers rely on accurate predictions to make informed management decisions and take appropriate actions [17]; commercial suppliers seek to understand how new seed varieties will perform in different areas [46]; governments and international organizations depend on early and accurate forecasts to anticipate disruptions in food security or import/exports [11].

In this work, we leverage an encoder-decoder framework to perform *in-field* prediction at the *pixel-level* and demonstrate superior performance, 5.41% MAPE test performance at the field level as seen in Figure 1. This approach not only provides exceptional performance at the field-level, but also enables the farmer to identify regions of his field which are under-performing and may benefit from further inspection and treatment. To do this we collect high-resolution (10cm/pixel) multi-spectral (RGB + NIR) imagery across 602 corn fields in the US-Midwest over 2 seasons (2020, 2021) and predict the final yield density at the *pixel-level* from imagery collected at the mid-way point of the season. We first baseline these approaches against traditional handcrafted approaches (which are still quite commonplace in agricultural and remote sensing works), and a tile-level

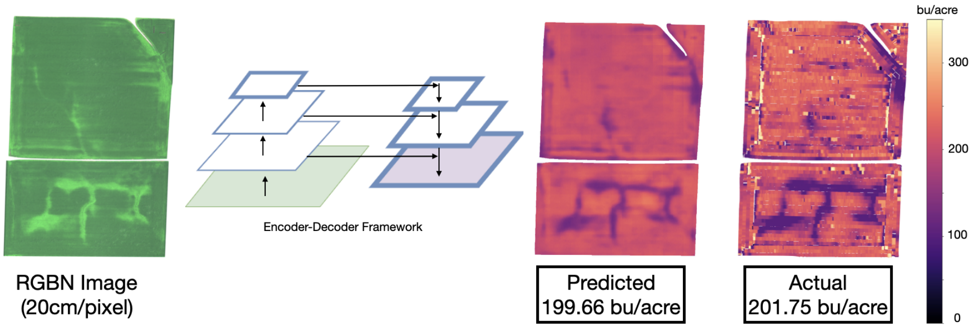

Figure 1. Our approach uses an encoder-decoder framework to directly predict the harvest *density* at the pixel-level from high-resolution RGB+NIR imagery. We demonstrate that this approach is superior to other methods which make predictions at coarser scales such as the tile-level.

CNN-based approach. Importantly, our dataset is unprecedented in size for field-level yield forecasting, our performance reaches SOTA-levels, and to the best of our knowledge, we are the first to demonstrate dense pixel-level yield forecasting from remote sensing imagery.

2. Related Work

Many existing works on crop yield forecasting focus on non-image based sources of information including soil, weather, and seed-variety. These sources of information may serve as inputs to mechanistic, i.e. simulation-based, models which are based on decades of research from agronomy and crop science [12, 30, 42]. Unfortunately, these methods often require complex data calibration to produce reasonable results and also suffer from huge computational overhead and lengthy run-times. Recently, many have looked to machine learning to complement or replace these approaches [48]. These include traditional machine learning algorithms [2, 21, 37, 43] as well deep learning approaches [8, 9, 24, 36]. Importantly, stand-count estimation, while a related task, is distinct from yield forecasting; to accurately predict the yield of a corn field requires information about the number of plants (i.e. stand-count) *and* the yield per plant. Recently, Khaki et al. [23] used a DNN to predict performance of corn hybrids throughout the United States from a dataset containing detailed hybrid information including genetic markers as well as environmental data such as weather [46]. Barbosa et al. [3] relied on machine data such as planting, spraying, and harvesting information, without any imagery data, to predict within-field yields.

As machine learning approaches have advanced rapidly over recent years, so too has remote sensing technology. Increased sensor resolution and channels beyond RGB have become more common thanks to improved satellites and collection via UAV and manned aircraft. However, much of the work done previously on yield forecasting from imagery is based on low-resolution remotely sensed data with predictions at a regional level [8, 45], or from imagery embedded on ground-based robots [27, 50]. Early approaches [5, 20] extracted hand-crafted features, often based on agriculturally-relevant vegetative indices like NDVI. With the success of deep learning approaches, more recent approaches have begun to apply neural networks for yield forecasting tasks. Leveraging a combination of deep Gaussian processes and long-short term memory, You et al. [52] predicted soybean yield at the county-level in the United States. While these analyses are important for anticipating issues surrounding food security, they are too coarse for providing individual farmers with actionable insight into his farm.

Within-field crop forecasting, specifically from remotely sensed data, is largely absent from past efforts. Nevavuori et al. [33] used imagery collected from UAVs combined with a 6-layer CNN to predict crop yield of different types across nine fields. While Barosa et al. [3] performed in-field predictions, that work was based on application data (e.g. planter and sprayer data) and did not use imagery. In both of these works, the authors used within-field cross-validation and other approaches to handle the very small number of fields (7) in their dataset; as no fields were fully held-out in the test set, it is unclear how such approaches would generalize to unseen fields as would be expected in real-world scenarios.

3. Data

3.1. Image Acquisition

We collected imagery data using manned aircraft flown over the Midwest US, primarily Illinois and Indiana, during the 2020 and 2021 growing seasons (April through September). The region was flown 13 times over this period to provide a longitudinal view of the crops' health and progression; flights were conducted roughly every two weeks. As the total area covered is quite expansive, covering over

5million acres, data acquisition is a nearly continuous process with different fields covered on different days for any given collection cycle. For this analysis, we focus on 603 fields, 402 in 2020 and 201 in 2021, in which corn was grown during that season and planter and harvester data was available (see Sec. 3.2).

Each image consisted of four channels: red, green, blue, and near-infrared captured at 10cm/pixel resolution. Orthorectification [4] and mosaicking were applied to the raw images to generate full-field imagery for analysis [13]. Notably, these full-field images are quite large, upwards of 1GB in size and 10k× 10k pixels in dimension. These images were also georeferenced to allow for subsequent alignment to equipment data.

3.2. Equipment Data

Modern planting (seeding) and harvesting equipment enables highly accurate tracking of seeding and harvesting rates. This equipment records the instantaneous velocity of the machine in addition to a seeding/harvesting rate at second to milliseconds intervals. These GPS guidance enabled tractors claim a positional accuracy range of 40cm to under 2cm with a real-time kinematic (RTK) positioning system [1]. We do not make any distinction as to the make and model of the equipment used on a given field, i.e., we do not make any explicit corrections for the specific precision of a given piece of equipment.

The final data output of these machines is a vector geodata file. The geodata file comprises one row (sample) per timestamp that includes the target rate and the applied rate, effectively providing a vector map of rates across the field. Speed is derived from the positional and timestamp data and is used to convert the raw output to a *density* in seeds/acre for the planter file and bushels/acre for the harvester. The vector map from the planter and harvest files were converted to a raster map by burning the applied rate value onto an empty raster grid to produce a file raster image with 20cm/pixel resolution and blurring with 5×5 normalized box filter. Because these are geofiles, they can be easily synced with the geo-referenced imagery collected in Sec. 3.1.

A boundary mask is constructed for each field to indicate which portion of the image is under active management, i.e. where planting has occurred and the grower expects to harvest crop. “Unmanaged” areas such as grassed waterways, houses, roads, etc. are excluded from analysis. This file is determined from the non-null areas generated by the harvest and planter files and is subsequently inspected for quality assurance.

All equipment data belongs to the farmers and has been used with permission: we have ensured the anonymity of the grower is protected by ensuring no figures, images or results disclose the identity of the grower.

3.3. Dataset

The current analysis focuses on predicting the end-of-season yield from imagery taken at or up to the middle of the season. Forecasting at different times during the season is the focus of future work. The *prediction flight* p was selected based on the Growing Degree Days (GDDs), aka. Growing Degree Units (GDUs), of the field. GDDs are used to estimate the growth and development of plants as development will only occur if the temperature exceeds some minimum development threshold, or base temperature (TBASE) determined experimentally for each crop; for corn TBASE is 50F [40]. The GDD value used in this work was obtained from the DarkSky API [19] based on the location of the field. For this work we select our prediction flight p as the first flight during the Pollination phase which corresponds to a GDD between 1135 and 1660, roughly mid-June through mid-July in these regions. Use of GDDs instead of flight date better normalizes the data to ensure prediction is made at roughly the same growth stage as fields may be planted over a wide range of dates (often over multiple months), and the plants' development is dictated by the local climate and seasonal weather conditions.

Images were downsampled using cubic resampling to produce images with 20cm/pixel resolution for analysis. These images were windowed into tiles of size 512×512 pixels with a stride of 512. Tiles containing $< 10\%$ data, were discarded. To ensure an even distribution of data in the train-validation-test sets, we applied stratified logic to the splitting as follows: fields for each season were grouped into five bins [100,150,200,225,250,300,350] based on their average yield per acre. Fields were then split by season-bin combination so that all tiles of a given field-season belonged to a single split (train, valid, test). This generates splits Train (2020): 18,859; Train (2021): 6,965; Valid (2020): 3,794; Valid (2021): 1,401; Test (2020): 3,652; and Test (2021): 1,463. For the majority of the experiments, splits for 2020 and 2021 were combined; Sec. 6.4 explores the out-of-season effect where the model is trained only on Train (2020) and performance is evaluated on Test (2021).

4. Hand-Crafted Models

As a baseline, we constructed a “tabular model” which leverages handcrafted features using approaches common to remote sensing and computational agriculture(Sec. 4.1) and traditional machine learning regression algorithms(Sec. 4.2).

4.1. Agronomic Feature Generation

Although deep learning approaches have proven to be tremendously successful in numerous domains, including remote sensing and computational agriculture, those two domains still feature a significant amount of work leveraging traditional computer vision, hand-crafted features, and non-deep learning approaches [14,29,49]; this due in part to their relatively strong performance on many tasks. Therefore, we compare approaches based on hand-crafted features to deep learning approaches. However, as the focus of this work is not on developing a model based on optimal hand-crafted features, we give a high-level overview of the process here and include additional details in the Supplementary Material.

Briefly, we first calculate the normalized difference vegetation index (NDVI) and green normalized difference vegetation index (GNDVI) across the field [31]. Next, we apply image processing including erosion, blurring, threshold, and connected components to identify anomalous regions from these NDVI and GNDVI maps. The field is then represented as $s = 4$ non-mutually exclusive binary masks F^s corresponding to the presence of agronomic features ("Ag-Feature") related to i) high stress, ii) low biomass, iii)low vigor, and iv)low (relative) growth; these are based on agronomic relationships derived from crop science and written in mutual collaboration with agronomists.

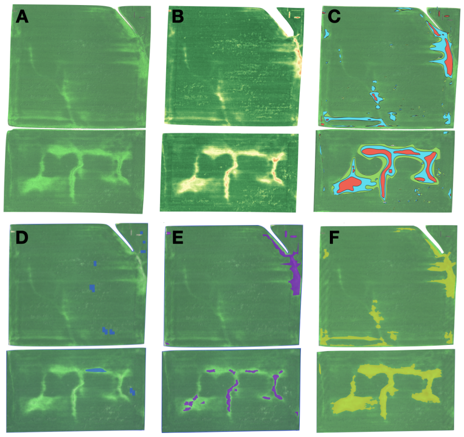

Figure 2. (A) The raw RGB+NIR is processed to generate different agronomic indicators like NDVI shown in (B). From these indices potentially problematic areas with different severity (red > cyan > green) are identified through anomaly detection and thresholding(C). With the aid of agronomists, features are constructed to identify areas of low growth (D), low vigor (E), high stress (F), and low biomass (none present in current example).

We run these feature-map creation steps for each flight image $I_1 : I_p$ up through the prediction flight p. Each Ag-Feature map is summed independently to generate a set of $s = 4$ final Ag-Feature maps

$$\mathbf{F}^s = \sum_{t=1}^{p} F_t^s \tag{1}$$

where an element of $\mathbf{F^s}$ is the number of times in the first p-flights of the season in which the s^{th} feature was present. These feature maps are tiled in accordance with the associated imagery (512 × 512) to enable per-tile feature construction; we then compute the mean, median, and max over each channel in $\mathbf{F^s}$ of that tile as features into the model.

Additionally, for each tile, we calculate features based on the mean, standard deviation, mean absolute deviation, standard deviation, and $[5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}]$ percentiles of common agronomic indices ("Index Features") such as NDVI, NDWI, SAVI, EVI, and GRNDVI; additional description of these indices is given in the Supplemental Material. We also use the mean, standard deviation, and skew of the red, green, blue, and NIR histograms of that tile. In certain experiments, we also directly used the latitude/longitude of the field and the exact GDD value of the prediction flight as features. Finally, from the planter file we extract the mean, standard deviation, and skew of the seeding rate distribution in that tile.

4.2. Models

We evaluate the performance of three common machine learning algorithms: Lasso, Random Forest (RF), and LightGBM regression. We explore different combinations of feature sources in Sec. 6.1 as well as the use of a feature-selection step based on minimum-redundancy-maximum-relevance (mRMR) to identify uncorrelated variables. In those models, all possible features are passed into the algorithm, and only the top subset are passed into the learning algorithm for training [41].

Models were fit to minimize the mean squared error (MSE) between the actual yield and predicted yield for that tile. That is, given the harvest map for a given tile which has a value in bushels(bu)/acre for each pixel, we define our loss as

$$MSE_{tile} = \sum_{i,j} Y_{ij} \circ M_{i,j} - \hat{Y}_{total} \tag{2}$$

where M_{ij} is the mask corresponding to the same area whose elements are 1 if the area is managed and 0 otherwise, and $\hat{Y}_{total}$ is the single value total yield predicted by the model.

All models were constructed using sklearn. Optimal hyperparameter values for each were found using the Scikit-Optimize package and are given in Supplementary Material.

5. CNN-Based Models for Tile and Pixel-Level Prediction

5.1. Input Representation

Each 512 × 512 tile is a 4-channel image consisting of the red, green, blue, and nir reflectance channels taken from the prediction flight I_p. RGBN channels were scaled by dividing by 2^{15}; this brings the naturally int16 values into the range 0-2, with the majority of the mass occurring in the range 0-1 because of the sensor's characteristics. We also explored the impact of using the Ag-Feature maps (Sec. 4.1) as additional input channels. Note that while only the prediction flight image is used, the Ag-Feature maps include some history about the field as they capture whether the field has ever experienced that feature; as each channel is the sum of occurrence through the p^{th} flight, this amount is divided by 10 to ensure the total was below 1. For one experiment we used the planter seeding rate map as an input; all values were divided by 50,000 to bring the resulting values near 1. The impact of NDVI and other indices was also explored; these by their nature are constrained to be in the range [-1,1].

5.2. Tile-level CNN

We first directly compare a deep-learning based approach to the hand-crafted tabular models of Sec. 6.1 by performing tile-level prediction. For each tile, we used only the four-channel RGBN image as the input to the model, and the total yield of the tile was predicted. We explored the impact of different architectures including VGG16 [44], ResNet-34,50 [15], RegnetY-040 [51], and Densenet-161 [18]. As done for the tabular models, we use MSE between the actual yield and predicted yield for that tile as given in Eq. (2).

5.3. Pixel-level CNN

Given the natural high-resolution of both the input (i.e. imagery) and target (i.e. harvester file), we next sought to perform *pixel-level* prediction to forecast the harvest at each point in the field directly. This approach recasts the problem of forecasting yield as a density-estimation task. That is, given an image X_{ij}, we predict the harvest density Y_{ij} in units/pixel. We can then calculate the total yield over a given region (e.g. tile, field, or other arbitrary region) as

$$TotalPredictedYield = \sum_{i,j} \hat{Y}_{ij} \circ M_{ij} \tag{3}$$

where M_{ij} is the mask corresponding to the same area whose elements are 1 if the area is managed and 0 otherwise.

We explored U-Net and FNP architectures with VGG16, ResNet-34, ResNet-50, RegnetY-040, and DenseNet-161 encoders. The loss was defined as the MSE between the actual yield density (in bushels/pixel) and the predicted yield density at each pixel in the tile.

$$MSE_{pixel} = \sum_{i,j} \|(Y_{ij} - \hat{Y}_{ij}) \circ M_{ij}\|_2^2 \tag{4}$$

5.4. Training

Encoders were initialized using Imagenet weights for the RGB channels. Weights in the first layer for the NIR channel and any additional input channels were initialized randomly. Horizontal and vertical flipping, transposition, and random rotation were used as augmentation during training.

Adam optimization with a learning rate of 1e-4 and weight decay of 1e-5 was used to minimize the loss. A multi-step learning rate scheduler with a multiplicative factor of $\gamma = 0.1$ to reduce the learning rate between the 5th and 15th epochs was used to control learning rate decay. The loss was defined as the MSE between the actual yield rate (i.e. bushels/acre) and predicted yield rate for each pixel in the tile. Models were trained with a batch size of 16 for 100 epochs with early stopping terminating with a patience of 10 epochs.

All CNN-based models were constructed in PyTorch for architecture construction and the Albumentations package [6] for augmentation. Models were trained on a machine with a single NVIDIA TitanRTX and Intel i9-9940X processor.

5.5. Metric Calculations

For the hand-crafted models of Sec. 6.1 and tile-level CNN of Sec. 5.2, the output of the model is the total yield of that tile. Field-level metrics are obtained by performing an aggregate over all the tiles of the field according to

$$AverageFieldValue = \frac{\Sigma(TileValue * TileArea)}{\Sigma(TileArea)} \tag{5}$$

where the tile area corresponds only to those areas in the tile which were planted (i.e. ignores intentionally unmanaged areas). Mean squared error (MSE), mean absolute error (MAE), and mean absolute percent error (MAPE) are then calculated on these totals.

For the pixel models of Sec. 5.3, the output of the model is in bushels/pixel where each pixel is 20cm^2 in area. Converting to bushels/acre at either the tile or field-level is achieved through simple dimensional analysis.

6. Experiments and Results

6.1. Tile-Level Regression

Both the Tabular and CNN-based models are used to perform tile-level regression. A sample result from the CNN model with ResNet-34 architecture is shown in Figure 3. The output of the model is total bushels for each tile; this is converted to bushels/acre for easy comparison across methods according to Sec. 5.5. We see that in addition to capturing the area of extremely low predicted yield on the far left, the model captures variations across the field.

For each model, MSE, MAE, and MAPE at the tile and field levels and report results in Table 1. All models, both the hand-crafted tabular models and CNN models, outperform the naive baseline by a significant margin. For all three traditional-ML algorithms, the best model included all features (raw image channels, agronomic indices, agronomic

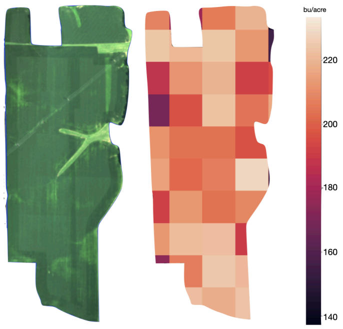

Figure 3. Sample result produced by the CNN-based tile-level model. Note that the bushels/acre of a given tile is over the "managed" area only; this is why the tiles containing the bright green area at the center-right of the figure (which corresponds to a grassed waterway), does not cause the prediction to be particularly low in that area. In contrast, the effect of the stress near the upper left of the image results in a below-average (purple) yield forecast for that tile.

features, planter data, and lat/long) passed through the feature selection step. Additionally, we show results for the best model versions using only features directly derivable from imagery. Note the best models include the planter data, which provides the number of seeds sewn in that tile, and has a significant impact; this suggests incorporation of this channel will be useful in future work. Results of additional experiments exploring the impact of different feature combinations are included in the Supplementary Material.

	Tile-Level: Test			Field-Level: Test		
	MSE	MAE	MAPE	MSE	MAE	MAPE
Naive Baseline	753.13	21.19	10.06	575.24	19.25	8.62
Lasso (best)	551.67	18.51	8.55	455.04	17.94	7.99
RF (best)	488.12	17.48	8.09	352.72	15.54	6.96
LGBM (best)	423.31	16.0	7.36	270.62	13.13	5.95
Lasso (image)	565.80	18.67	8.61	459.37	17.59	7.84
RF (image)	513.61	17.71	8.18	408.83	16.44	7.36
LGBM (image)	515.39	17.67	8.10	435.60	16.41	7.35
VGG16	511.46	17.78	7.96	378.28	15.45	6.85
ResNet-34	428.64	16.45	7.49	281.53	13.69	6.16
ResNet-50	389.24	15.58	7.06	251.57	12.99	5.78
RegnetY-040	388.48	15.75	7.12	251.39	13.08	5.82
Densenet-161	533.10	17.98	8.06	390.05	14.78	6.38

Table 1. Performance of Tile-Level Regression Models

All CNN-based tile-level models here use RGBN imagery only. Every CNN model out-performed every hand-crafted model in Table 1 with the exception of the best-LGBM model which used features from multiple sources (e.g. planter, lat/long) in addition to imagery. The best CNN architecture, ResNet-50, produced results with 1% MAPE better than the best LGBM image-only model. While the hand-crafted features performed (perhaps surprisingly) well, the CNN still achieves better performance when provided equivalent input information (i.e. image only). However, the performance of the multi-source best hand-crafted models suggest there is still significant opportunity for incorporating this information into the CNN approaches, which is the focus of future work.

6.2. Pixel-Level Regression

Our pixel-level regression approach directly predicts the harvest map from the input imagery. We used two common encoder-decoder frameworks, U-net [38] and Feature Pyramid Network(FPN) [26] and explored a combination of architectures and backbones as described in Sec. 5.3. Results are shown in Table 2 and residual analysis is provided in the Supplementary Material.

Every pixel-level model (Table 2) outperforms every hand-crafted image-only model (Table 1 Middle) except for FPN VGG16. Furthermore, every pixel-level U-Net CNN model except RegnetY-040 matches or outperforms its tile-level counterpart (Table 1 Bottom). Pixel-level FPN Densenet-161 significantly outperforms all other models. Since the same information is being extracted by the same encoder, this suggests that the dense loss signal afforded to the model by predicting the pixel-level harvest file directly as a density map has tremendous benefit.

	Tile-Level: Test			Field-Level: Test		
	MSE	MAE	MAPE	MSE	MAE	MAPE
U-Net VGG16	504.82	17.46	7.82	392.38	15.96	7.06
U-Net ResNet-34	395.24	15.61	7.03	274.49	13.26	5.91
U-Net ResNet-50	365.36	15.19	6.88	296.47	13.42	5.98
U-Net RegnetY-040	394.96	15.72	7.11	256.30	13.12	5.89
U-Net DenseNet-161	379.54	15.42	6.98	255.23	12.88	5.78
FPN VGG16	525.50	17.74	7.83	436.88	12.24	7.09
FPN ResNet-34	371.26	15.18	6.84	269.58	13.33	5.91
FPN ResNet-50	395.23	15.68	7.06	264.34	12.92	5.73
FPN RegnetY-040	421.49	16.38	6.62	261.72	13.04	5.80
FPN DenseNet-161	347.01	14.77	6.59	234.97	12.29	5.41

Table 2. Performance of Pixel-Level Regression Models

Sample output of the U-Net architecture with DenseNet-161 model is shown in Fig. 4. This figure highlights two "good" samples at the top and two "bad" samples at the bottom. We see that even in the bad examples, the model does a good job identifying struggling areas on the field. This is not surprising as we saw that even the models based on hand-crafted features derived from agronomic indices like NDVI managed to identify these areas. However, this dense pixel-level model does a significantly better job determining the magnitude of the effect as seen by the overall tile and field-level performance. Furthermore, we see that appearance of fields can vary dramatically with healthy "green" crops covering many different shades due to lighting conditions as well as seed variety; the pixel-level model captures these variations without the significant burden of crafting detailed hand-crafted features to address these different

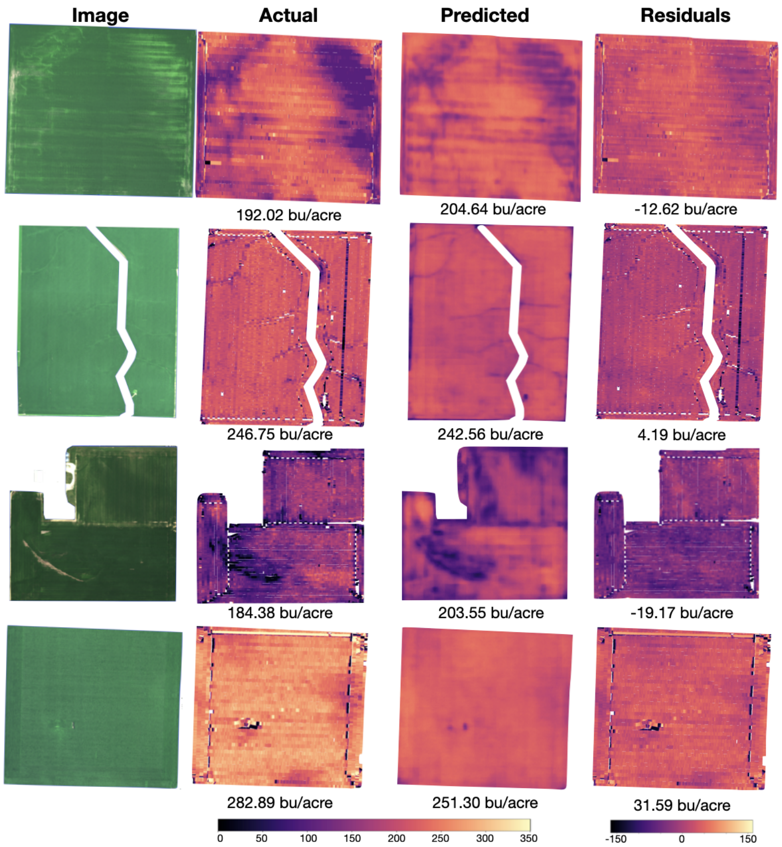

Figure 4. Field image, actual harvest map, predicted harvest density, and residuals from two "good" fields (top two rows) and two "bad" examples (bottom two rows).

sources of variability.

6.3. Impact of Alternate Input Representation

Including additional channels based on agronomic indices make the input representation overcomplete. Inclusion of outlier feature channels, however, provides additional information as they contains information about whether that feature was present in *previous* flights through the sum.

For this set of experiments, we focused on the U-Net architecture with DenseNet-161 backbone. Any additional features were simply added as additional input channels beyond the original RGBN. Results in Table 3 show the impact of these additional channels. The overcomplete representation with the addition of the NDWI and SAVI channels did not improve performance; this suggests the network is learning sufficiently expressive features that do not benefit from explicit incorporation of hand-crafted features which have been so prominent in remote sensing work.

In stark contrast to the results seen when adding the planter file to the tabular model, inclusion of the planter file here did not improve results. In this experiment, the planter file was included only an additional input channel. However, incorporation into the network through late-fusion techniques as in [3] may prove more effective; fusion of additional channels is the focus of ongoing work.

Only the incorporation of the stress feature improved results; this is perhaps not surprising as it incorporates additional information from previous flights and exploratory analysis demonstrated the highest correlation between the stress feature and yield among the outlier features. Incorporating all outlier features (including stress), however, caused the performance to degrade slightly. Nevertheless, the performance boost from the stress feature suggests that incorporation of earlier flights directly, could improve results even further.

6.4. Year-Over-Year Domain Shift

Results presented thus far have been trained and evaluated on data combined from 2020 and 2021. However, in reality, such a model would never have access to harvest results from the next season. To understand this effect, we trained a U-Net with DenseNet-161 encoder only on the 2020 train set and evaluated performance both on

	Tile-Level:Test			Field-Level:Test		
UNet DenseNet-161	MSE	MAE	MAPE	MSE	MAE	MAPE
RGBN	379.54	15.42	6.98	255.23	12.88	5.78
RGBN+NDWI+SAVI	389.64	15.50	7.04	263.61	13.09	5.89
RGBN+Planter	381.76	15.30	6.93	268.61	13.32	5.98
RGBN+Stress	356.21	14.90	6.72	234.89	12.06	5.40
RGBN+All Outliers	394.99	15.68	7.11	267.32	13.42	6.04

Table 3. Impact of additional feature channels

	Tile-Level			Field-Level		
	MSE	MAE	MAPE	MSE	MAE	MAPE
2020 Val	503.47	16.71	6.67	305.08	12.62	5.99
2020 Test	370.19	15.10	6.75	218.33	11.98	5.41
2021 Val	611.21	20.28	9.19	384.09	16.74	7.52
2021 Test	718.17	21.61	9.88	601.52	20.37	8.71

Table 4. Out-of-Domain Analysis

the in-domain (2020) validation and test sets, as well as the out-of-domain (2021) validation and test sets. Results in Sec. 6.4 show that while the in-domain 2020 results remain strong (even though the training data is less), the out of domain 2021 test and validation performance decreases significantly. This is not surprising as there is known upward annual data drift in corn harvest in the US [25]. Fortunately, there are ways which this can be addressed that are explored in the Discussion.

7. Discussion

This is the first work to explore yield-forecasting as a density-estimation problem from remote sensing imagery to enable in-field (i.e. pixel-level) yield prediction. Our approach simultaneously produces improved results at the field-level and also provides in-field predictions which growers can use to identify struggling areas and make important management decisions.

Advances in remote sensing, computer vision, and smart farming equipment are enabling remarkable opportunities for precision agriculture. As these technologies continue to improve, so too will our ability to forecast yield. While the current work featured imagery collected from manned aircraft, the approach is agnostic to data source. High-resolution satellite technology is rapidly improving in both quality and coverage area and could easily be used an image source for this work; satellite in particular will become important for extending the application of yield forecasting globally, especially to developing countries which may benefit from it the most.

In the current work we explored yield forecasting only from a single point in time. Future work will explore the use spatiotemporal modeling to incorporate the field's progression throughout the season leveraging all collected imagery. Use of additional imagery is expected to not only produce improved predictions at mid-season, but also to enable sound predictions *earlier* in the season. Incorporation of other modalities such as soil, topography, and weather are also expected to boost performance. Influence of regional and annual trends will also be explored in future work. Finally, extending this analysis to multiple crops is of key interest and central to addressing issues around global food security.

We saw the effect of out-of-season domain shift on the model's performance in Sec. 6.4. This is not surprising as weather and seed selection are two of the most influential factors on a field's performance and these can vary dramatically year-over-year. However, recent work using remote sensing and decades of yield data at the county or regional level have demonstrated strong results in forecasting regional trends very early into an unseen season. Fusion of these low-resolution temporal models, with the very high-resolution models explored here would likely improve out-of-domain performance significantly. Direct incorporation of constraints imposed by known agronomic principals is also an interesting area of exploration which could enable greater generalization.

A significant benefit of our approach is in fact its straightforwardness and simplicity. While the success of deep learning approaches is not a surprise to the computer vision community, it is important to reiterate here because of the initially slow adoption of deep learning within agriculture and remote sensing. Several of the handcrafted models performed surprisingly well; solid performance from these types of approaches is a key reason they continue to be commonplace in remote sensing and computational agriculture. However, feature generation for the handcrafted tabular model is a painstaking task which requires numerous steps involving image processing, statistics, and incorporation of agricultural domain knowledge; parameters are largely picked based on expert evaluation or knowledge from research in agronomy or crop science. Exhaustively searching for the best combination of features and image processing parameters is impossible and also does not generalize to other tasks or scenarios. And while individual steps of the processing can be articulated, certain design choices or parameter values may appear arbitrary, offering little to no clarity or actual interpretability to the end consumer of the model's output. Furthermore, small changes to the sensors or image source may render the current set of parameter choices invalid and the ability to generalize to different crops or soil types is severely limited. In contrast, the deep learning approaches produce significantly improved results and provide a clear path forward to adaptation, generalization, and model improvement. We hope that this and other work using deep learning for remote sensing and precision agriculture continues to fuel adoption in these domains.

References

[1] John deere guidance systems. Technical report, John Deere, 2021. 3

[2] AT M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, and Rashedur M Rahman. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh. In *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 1–6. IEEE, 2015. 2

[3] Alexandre Barbosa, Rodrigo Trevisan, Naira Hovakimyan, and Nicolas F. Martin. Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*, 170:105197, 2020. 1, 2, 7

[4] Oscar Rosario Belfiore and Claudio Parente. Orthorectification and pan-sharpening of worldview-2 satellite imagery to produce high resolution coloured ortho-photos. *Modern Applied Science*, 9:122–130, 08 2015. 3

[5] Douglas K Bolton and Mark A Friedl. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013. 2

[6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5

[7] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020. 1

[8] Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, 151:61–69, 2018. 2

[9] Andrew Crane-Droesch. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13(11):114003, 2018. 2

[10] Saba Dadsetan, Gisele Rose, Naira Hovakimyan, and Jennifer Hobbs. Detection and prediction of nutrient deficiency stress using longitudinal aerial imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14729–14738, May 2021. 1

[11] Felix Dodds and Jamie Bartram. *The water, food, energy and climate Nexus: Challenges and an agenda for action.* Routledge, 2016. 1

[12] Scott T Drummond, Kenneth A Sudduth, Anupam Joshi, Stuart J Birrell, and Newell R Kitchen. Statistical and neural methods for site–specific yield prediction. *Transactions of the ASAE*, 46(1):5, 2003. 2

[13] Feng Gao, Jeffrey G Masek, and Robert E Wolfe. Automated registration and orthorectification package for landsat and landsat-like data processing. *Journal of Applied Remote Sensing*, 3(1):033515, 2009. 3

[14] Liang Han, Guijun Yang, Huayang Dai, Bo Xu, Hao Yang, Haikuan Feng, Zhenhai Li, and Xiaodong Yang. Modeling maize above-ground biomass based on machine learning approaches using uav remote-sensing data. *Plant methods*, 15(1):1–19, 2019. 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Jennifer Hobbs, Prajwal Prakash, Robert Paull, Harutyun Hovhannisyan, Bernard Markowicz, and Greg Rose. Large-scale counting and localization of pineapple inflorescence through deep density-estimation. *Frontiers in Plant Science*, 11:2157, 2021. 1

[17] T Horie, M Yajima, and H Nakagawa. Yield forecasting. *Agricultural systems*, 40(1-3):211–236, 1992. 1

[18] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 5

[19] Apple Inc. Dark sky api. 3

[20] David M Johnson. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the united states. *Remote Sensing of Environment*, 141:116–128, 2014. 2

[21] Elisa Kamir, François Waldner, and Zvi Hochman. Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160:124–135, 2020. 2

[22] Keren Kapach, Ehud Barnea, Rotem Mairon, Yael Edan, and Ohad Ben-Shahar. Computer vision for fruit harvesting robots–state of the art and challenges ahead. *International Journal of Computational Vision and Robotics*, 3(1-2):4–34, 2012. 1

[23] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019. 1, 2

[24] Saeed Khaki, Lizhi Wang, and Sotirios V. Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10, 2020. 2

[25] Christopher J Kucharik and Navin Ramankutty. Trends and variability in us corn yields over the twentieth century. *Earth Interactions*, 9(1):1–29, 2005. 8

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 6

[27] Simon Madec, Xiuliang Jin, Hao Lu, Benoit De Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Frédéric Baret. Ear density estimation from high resolution rgb imagery using deep learning technique. *Agricultural and Forest Meteorology*, 264:225–234, 2019. 2

[28] Lonesome Malambo, Sorin Popescu, Nian-Wei Ku, William Rooney, Tan Zhou, and Samuel Moore. A deep learning semantic segmentation-based approach for field-level sorghum panicle counting. *Remote Sensing*, 11(24):2939, 2019. 1

[29] Aaron E Maxwell, Timothy A Warner, and Fang Fang. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018. 4

[30] Robert L McCown, Graeme L Hammer, John Norman Gresham Hargreaves, Dean P Holzworth, and David M Freebairn. Apsim: a novel software system for model development, model testing and simulation in agricultural systems research. *Agricultural systems*, 50(3):255–271, 1996. 2

[31] SM Moges, WR Raun, RW Mullen, KW Freeman, GV Johnson, and JB Solie. Evaluation of green, red, and near infrared bands for predicting winter wheat biomass, nitrogen uptake, and final grain yield. *Journal of plant nutrition*, 27(8):1431–1441, 2005. 4

[32] United Nations. World could face water availability shortfall by 2030 if current trends continue, secretary-general warns at meeting of high-level panel. 1

[33] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and electronics in agriculture*, 163:104859, 2019. 2

[34] United Nations Department of Economic and Social Affairs. 1

[35] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):1–12, 2019. 1

[36] T Venkat Narayana Rao and S Manasa. Artificial neural networks for soil quality and crop yield prediction using machine learning. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 5(1):57–60, 2019. 2

[37] José R Romero, Pablo F Roncallo, Pavan C Akkiraju, Ignacio Ponzoni, Viviana C Echenique, and Jessica A Carballido. Using classification algorithms for predicting durum wheat yield in the province of buenos aires. *Computers and electronics in agriculture*, 96:173–179, 2013. 2

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[39] Inkyu Sa, Zetao Chen, Marija Popović, Raghav Khanna, Frank Liebisch, Juan Nieto, and Roland Siegwart. weednet: Dense semantic weed classification using multispectral images and mav for smart farming. *IEEE Robotics and Automation Letters*, 3(1):588–595, 2017. 1

[40] EC Schneider and SC Gupta. Corn emergence as influenced by soil temperature, matric potential, and aggregate size distribution. *Soil Science Society of America Journal*, 49(2):415–422, 1985. 3

[41] Ram Seshadri. featurwiz. 4

[42] Mohsen Shahhosseini, Rafael A Martinez-Feria, Guiping Hu, and Sotirios V Archontoulis. Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, 14(12):124026, 2019. 2

[43] Avat Shekoofa, Yahya Emam, Navid Shekoufa, Mansour Ebrahimi, and Esmaeil Ebrahimie. Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PloS one*, 9(5):e97288, 2014. 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[45] Jie Sun, Liping Di, Ziheng Sun, Yonglin Shen, and Zulong Lai. County-level soybean yield prediction using deep cnn-lstm model. *Sensors*, 19(20):4363, 2019. 2

[46] Syngenta. Crop challenge in analytics". 1, 2

[47] Gabriel Tseng, Hannah Kerner, Catherine Nakalembe, and Inbal Becker-Reshef. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1111–1120, 2021. 1

[48] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020. 2

[49] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020. 4

[50] Qi Wang, Stephen Nuske, Marcel Bergerman, and Sanjiv Singh. Automated crop yield estimation for apple orchards. In *Experimental robotics*, pages 745–758. Springer, 2013. 2

[51] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification. *arXiv preprint arXiv:2101.00590*, 2021. 5

[52] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*, 2017. 1, 2