Monitoring Ethiopian Wheat Fungus with Satellite Imagery and Deep Feature Learning

Reid Pryzant*, Stefano Ermon*, and David Lobell**
*Department of Computer Science
**Department of Earth System Science
Stanford University
{rpryzant,ermon,dlobell}@stanford.edu

Abstract

Wheat is the most important Ethiopian crop, and rust one of its greatest antagonists. There is a need for cheap and scalable rust monitoring in the developing world, but existing methods employ costly data collection techniques. We introduce a scalable, accurate, and inexpensive method for tracking outbreaks with publicly available remote sensing data. Our approach improves existing techniques in two ways. First, we forgo the spectral features employed by the remote sensing community in favor of automatically learned features generated by Convolutional and Long Short-Term Memory Networks. Second, we aggregate data into larger geospatial regions. We evaluate our approach on nine years of agricultural outcomes, show that it outperforms competing techniques, and demonstrate its predictive foresight. This is a promising new direction in crop disease monitoring, one that has the potential to grow more powerful with time.

1. Introduction

Ethiopia is the largest sub-Saharan cultivator of wheat [7]. Three million tons of the cereal are harvested each year from 1.6 million ha of rain-fed land. The grain is a staple for more than five million households and an important fixture of the region's culinary traditions.

Wheat rusts, primarily *P. graminis* (stem rust) and *P. striiformis* (stripe rust), but also *P. dispersa* (leaf rust) are heteroecious fungi of the *Puccinia* genus and important biotic constraints to Ethiopian wheat production [34]. Repeated and explosive rust epidemics have swept the nation. Most recently, a 2014 outbreak of the stem rust *TTKSK* resulted in yield losses in excesses of 85% in some regions [27]. It would be prudent to explore cheap, rapid, and accurate methods for the identification and assessment of these pathogens so that farmers and governments can better inform their macro-scale management practices.

Figure 1. Fungal infections of interest. From left to right: *P. dispersa* (leaf rust), *P. graminis* (stem rust), *P. striiformis* (stripe rust).

Wheat rust symptoms include a gradual drying of the host's photosynthetic apparatus. Infected individuals loose both their pigmentation and photosynthetic ability [1, 30]. These observations suggest that remote sensing, a globally available and economically viable source of data, may be used to monitor the disease. These multispectral images capture wavelengths beyond the visible spectrum, have high spatial and temporal resolution, and thus contain a vast wealth of information on plant health through time.

Remote sensing is frequently used in computational sustainability studies, including species distribution modeling [8, 20], poverty mapping [37, 17], and even the monitoring of wheat diseases like powdery mildew [39] and rust [35, 1]. However, prior work on crop disease monitoring has not produced methods which are cheap and scalable, especially for the developing world. The majority of work within this domain has focused on ground- and air-level spectrography, which necessitates sophisticated and expensive data collection techniques. Furthermore, these data are high-dimensional and unstructured, so useful features that are predictive of agricultural outcomes are hard to extract. Previous work relies on painstakingly handcrafted features which may overlook latent structure in these data.

In this paper, we propose a novel technique for automatic wheat rust monitoring from satellite imagery. This is a difficult and interesting task, as it involves detecting a single

species of fungus from high-dimensional satellite imagery covering large swaths of terrain. In lieu of an expensive experimental study, we use data from a cheap and easily scalable observational survey in Ethiopia. Our approach is inspired by progress in deep feature learning and the benefits it has imparted on computer vision [23, 19]. Experimental results indicate that automatically learned features outperform traditional spectral features by a wide margin of 0.09 AUC, or 16%. Furthermore, our results suggest that this dichotomy in efficacy will grow as surveys of this kind continue. Our results, then, serve as a valuable proof-of-concept for a new direction in crop disease monitoring, one that is especially scalable in developing world and will grow more powerful in the future.

2. Related Work

Remote sensing data has been used to monitor wheat rust [35, 39, 1]. However, all existing approaches we are aware of rely on one or all of (1) ground truth data from experimental plots that were specially prepared or identified for the study (2) hand-crafted features, and (3) lower-order moments of sensing data (*e.g.* "mean", the first moment). All of these aspects have suboptimal qualities that inhibit their utility, especially in developing countries like Ethiopia.

Prior work often relies on data from elaborate experimental studies. For example, Liu *et al.* [24] used agricultural records to select 90 wheat plots and collected sensing data from those plots. Moshou *et al.* [26] sowed 12 plots of wheat, inoculated half of them with rust, and used a ground-level spectrometer to measure reflectance throughout the growing season. These sophisticated experimental designs are prohibitively expensive for the developing world. Furthermore, it is unlikely that models trained on these carefully regulated conditions would be usefully applicable to heterogeneous wheat fields in a polyculture.

All existing work we are aware of relies on hand-crafted features. Some widely used features include Triangular Vegetation Index (TVI) [13, 35], Normalized Difference Vegetation Index (NVDI) [28, 29, 1], and the Chlorophyll Absorption in Reflectance Index (CARI) [36, 5]. These existing features are relatively crude indices which depend on a small number of bands (usually two). We are the first to predict similar outputs with features that are automatically learned from the data. Our experimental results indicate that these learned features are both more effective and more diverse.

All prior work calculates predictive features from the pixels of remote sensing imagery [39, 15]. This means that prior work provides features for each possible location within the spaces they are interested in. Higher-order moments of the features are rarely explored. In contrast, our model learns tractable features from the *entire distribution* of pixels for a region of interest.

3. Deep Learning Models

We continue by reviewing the various building blocks of our deep learning model, which we use to automatically learn features. We make use of Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs). These models are equivalent to a complex nonlinear mapping from inputs to outputs. The mappings are stacked such that one layer's output is another layer's input. Such stacked architectures often learn successive representations of the data, distilling it into a more structured form that is easier to classify.

3.1. Deep Neural Networks

DNNs are the basic form of feed-forward neural network. They are composed of a series of fully connected layers, where each layer takes on the form

$$\boldsymbol{y} = f(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

Note that $\boldsymbol{x} \in \mathbb{R}^n$ is a vector of inputs (e.g. from a previous layer), $\boldsymbol{W} \in \mathbb{R}^{y \times n}$ is a matrix of parameters, $\boldsymbol{b} \in \mathbb{R}^y$ is a vector of biases, $\boldsymbol{y} \in \mathbb{R}^y$ is an output vector, and $f(\cdot)$ is some nonlinear activation function, *e.g.* the sigmoid $\sigma(x) = 1/1 + e^{-x}$.

3.2. Convolutional Neural Networks

CNNs are a breed of deep neural network that perform exceptionally well on data with a grid-like topology. CNNs are DNNs that employ convolutions instead of matrix multiplication in at least one layer.

A convolution operation involves a *filter* $\boldsymbol{w} \in R^{h \times h'}$ which is dragged around the input and, at each step, applied to whatever patch it is residing in with element-wise multiplication and sum reduction. For example, let $\boldsymbol{x} \in \mathbb{R}^{z \times z'}$ be a matrix of inputs. Then feature c_i is generated by

$$c_i = f\left(\sum_k \sum_j (\boldsymbol{w} \odot \boldsymbol{x}_{i:i+h-1,i:i+h'-1})_{j,k} + b\right) \tag{2}$$

where f is a non-linear activation function and $b \in \mathbb{R}$ is a bias term. Intuitively, filters are detectors for patterns in the input. Output features reflect the "best match" for those patterns in the input.

3.3. Long Short-Term Memory Networks

Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are effective tools for learning structure from sequential data [6]. RNNs take a vector $\boldsymbol{x}_t$ at each timestep. They compute a hidden state vector $\boldsymbol{h}_t$ at each timestep by applying nonlinear maps to the the previous hidden state $\boldsymbol{h}_{t-1}$ and the current input $\boldsymbol{x}_t$ (note that h_0 is initialized to $\vec{0}$):

$$\boldsymbol{h}_t = \sigma\Big(\boldsymbol{W}^{(hx)}\boldsymbol{x}_t + \boldsymbol{W}^{(hh)}\boldsymbol{h}_{t-1}]\Big) \tag{3}$$

Though RNNs can in theory model dependencies of indefinite length, training them is difficult because repeated applications of $\sigma(\cdot)$ drives error signals to exponential decay as they propagate back through time.

Long Short-Term Memory Networks (LSTMs) are a variant of the above RNN formulation. LSTMs can more effectively model long-term temporal dependencies by coping with the vanishing gradient problem inherent in their predecessor [14, 4, 12]. LSTMs behave much like RNNs but provide finer control over the mixing of past and present hidden states.

LSTMs have a pair of vectors that can remember information: $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$. First, $\boldsymbol{c}_t$, the *memory cell*, is a blending of the previous timestep's memory cell and a candidate cell $\widetilde{\boldsymbol{c}}$ that is proposed by the model:

$$\widetilde{\boldsymbol{c}}_t = \tanh\left(\boldsymbol{W}^{(c)}\boldsymbol{x}_t + \boldsymbol{U}^{(c)}\boldsymbol{h}_{t-1}\right) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \circ \widetilde{\boldsymbol{c}}_t \tag{5}$$

Note that the mixing in equation (3) is controlled by the *input gate* $\boldsymbol{i}_t$ and *forget gate* $\boldsymbol{f}_t$, both a function of the input and past hidden state:

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}^{(i)}\boldsymbol{x}_t + \boldsymbol{U}^{(i)}\boldsymbol{h}_{t-1}\right) \tag{6}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}^{(f)}\boldsymbol{x}_t + \boldsymbol{U}^{(f)}\boldsymbol{h}_{t-1}\right) \tag{7}$$

Next, the second memory vector (the new *hidden state*) is computed by throttling the new memory cell, with the degree of strangulation determined by an *output gate* $\boldsymbol{o}_t$:

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}^{(o)}\boldsymbol{x}_t + \boldsymbol{U}^{(o)}\boldsymbol{h}_{t-1}\right) \tag{8}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{c}_t) \tag{9}$$

Note that the $\boldsymbol{U}$ and $\boldsymbol{W}$ matrices are parameters to be learned during training. Predictions can be generated from these networks by attaching fully connected layers to each cell's hidden state vector.

4. Data Inputs

In this section, we describe the raw inputs to the proposed technique.

We obtained 8,554 field-level observations from an ongoing survey organized by the International Maize and Wheat Improvement Center (CIMMYT). Each observation consists of a date, location (latitude, longitude), and severity ratings for stem, stripe, and leaf rust on 0 to 3 point scales. These observations span nine growing seasons (2007 - 2016). We excluded the 2009 season due to a surfeit of incomplete observations, leaving us with 7,678 field-level observations.

We also obtained remote sensing data on surface reflectance (bands 1-7, MOD09A1), land surface temperature (bands 1 & 5, MYD11A2), and gross primary productivity (GPP, band 1, MYD17A2H). These data come from the Moderate Resolution Imaging Spectroradiometer (MODIS) apparatus aboard NASA's Terra and Aqua satellites [18]. This imagery has a 500 meter spatial resolution, an 8-day temporal resolution, and it covers the entire duration of our survey data. We used the IIASA-IFPRI global cropland map to mask out pixels not corresponding to farmland [9], but did *not* explicitly identify pixels belonging to wheat fields.

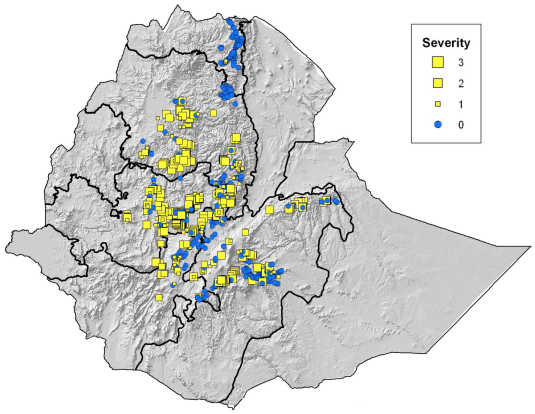

Figure 2. An illustrative example of our CIMMYT survey data. A subset of observations for the 2010 growing season are displayed.

5. Rust Monitoring as a Prediction Task

There is a strong need for cheap, scalable crop disease monitoring. However, but as far as these authors are aware, there is no precedent for this line of work. As such, we executed a series of novel pre-processing steps to fit this problem into the framework of a tractable supervised learning task.

5.1. Reconciling Survey and Sensing Data

Field-level observations are made with pinpoint geospatial accuracy, whereas satellite imagery covers broad swaths of terrain. We reconciled these two data sources by aggregating survey data into larger geographical regions and embedding our imagery in a lower dimensional space.

MODIS imagery has an appropriately fine-grained temporal resolution (8 days) but relatively course spatial resolution (500 $\times$ 500 meters). A single MODIS pixel may engulf several of the fields that CIMMYT surveyors made observations on. Because of this, we binned our CIMMYT data into larger geographic regions (Ethiopian Woreda's, roughly equivalent to American counties). This left us with 884 final data points, where each point maps to some region in a particular season.

Motivated by the scarcity of data, we next executed a

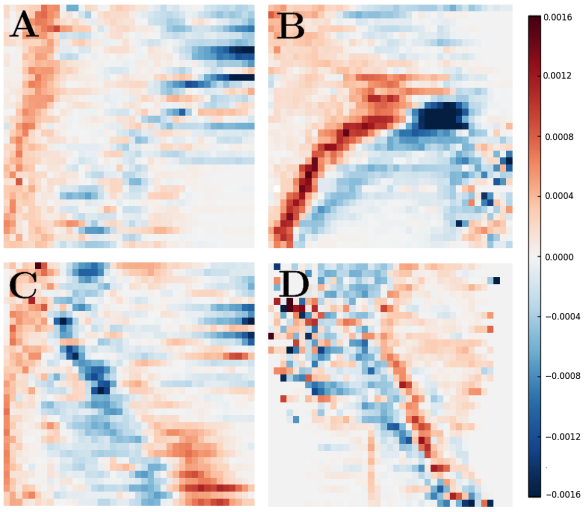

Figure 3. Differences between mean histogram sequences for regions with and without outbreaks. Each row i corresponds to $\Delta \boldsymbol{H}^{(i)} = \mu^{neg}_{\boldsymbol{H}^{(i)}} - \mu^{pos}_{\boldsymbol{H}^{(i)}}$, so positive (red) cells indicate an imbalance towards outbreak-free regions, while negative (blue) cells indicate the opposite. Time is moving down the y-axis; the top row depicts $\Delta \boldsymbol{H}^{(1)}$, and the last, $\Delta \boldsymbol{H}^{(m)}$. Red (a), GPP (b), mid-infared (c), and daytime temperature (d) are depicted. It is apparent that there are minute differences between positive and negative histograms. For example, note the trend towards cooler temperatures and increased bioactivity in fungus-prone regions.

variant of the dimensionality reduction procedure proposed by You *et al*. [38]. This procedure makes the assumption of permutation invariance; we don't expect the location of each pixel to be a strong predictor of rust outbreaks at the regional level. Rather, it is the shifting landscape of hues and intensities that should be considered. This assumption forces the model to eschew some location-specific information (*i.e*. soil properties, elevation), but we believe it affords tractability because we can consider low-dimensional *pixel intensity distributions* instead of high-dimensional images.

For each image i and band j, we discretized the pixels of $\boldsymbol{I}^{(i)}_j$ into w bins, then computed the histogram of pixel intensities $\boldsymbol{H}^{(i)}_j$. We next applied a novel centering technique to the data. We considered the complete timeseries of histograms for each region, season, and band, then selected individual buckets from these sequences and computed summary statistics of their intensities through time. Last, we subtracted and divided these buckets by their respective means and standard deviations. In this way, the bands of each histogram timeseries were independently standardized across time *and* bucket.

5.2. Label Assignment

We selected a binary labeling scheme that labels each point according to whether, at any point during its corresponding season, there was a significant outbreak of rust. These labels were determined with the following heuristic:

$$y^{(p)} = \mathbf{1}\left[1 < \sum_{k=0}^{n} \frac{m(\boldsymbol{x}^{(p)}_k)}{d(\boldsymbol{x}^{(p)}_k)}\right] \tag{10}$$

Where $y^{(p)}$ is the label of point p, $\boldsymbol{x}^{(p)}_k$ is the k^{th} field-level observation binned into point p, $m(\mathbf{x})$ is the max of $\mathbf{x}$'s three severity ratings, and $d(\boldsymbol{x})$ is the number of days $\boldsymbol{x}$ is away from the *end* of the growing season. Intuitively, this heuristic assigns positive labels to regions in seasons where, on average, field-level observations indicated infections, with observations on more mature plants given more importance. This temporal downweighting was motivated by a desire to reduce false negatives early on in the season. Observers are less likely to observe rust early in the season because wheat fields haven't matured to the point where they are viable hosts for the pathogen.

5.3. Prediction Task

In all, for each region and season, we have a sequence of image histograms that we would like to map to binary labels. Formally, this is a set

$$D = \Big\{ \Big((\boldsymbol{H}^{(1)}, \cdots, \boldsymbol{H}^{(m)}, g_{loc}, g_{year}), y \Big), \tag{11}$$

$$\cdots$$

$$\Big((\boldsymbol{H}^{(1)}, \cdots, \boldsymbol{H}^{(m)}, g_{loc}, g_{year}), y \Big) \Big\}$$

of histogram sequences $\boldsymbol{H}^{(1)}, \cdots, \boldsymbol{H}^{(m)}$, region identifiers g_{loc}, growing seasons g_{year} and corresponding outbreak indicators y. Note that each image histogram $\boldsymbol{H}^{(k)}$ consists of the 10 aforementioned bands $\boldsymbol{H}^{(k)}_1, \cdots, \boldsymbol{H}^{(k)}_{10}$.

Our objective is to map each histogram timeseries to the appropriate labels. This is equivalent to automatically identifying wheat rust outbreaks from satellite imagery and sparse field-level observations, which has never been done before. We will also consider the harder problem of predicting from subsequences $\boldsymbol{H}^{(1)}, \cdots, \boldsymbol{H}^{(r)}$ with $r < m$. This is equivalent to forecasting outbreaks well before the harvest date, when only a subset of the sensed data is available.

6. Experiments

In this section we demonstrate (1) the viability of the proposed technique and (2) how, within this domain, automatically learned features are substantially more powerful than traditional spectral indices.

We conducted four sets of experiments. The first is a comparative evaluation between the proposed method and a classifier with traditional spectral features. The second

explores the predictive potential of the proposed method. The third elucidates how much information learned features can draw from each band, and the fourth investigates the impact of continued data collection.

6.1. Experimental Configuration

Our experiments were conducted with a deep learning model that was designed for the task at hand (Figure 4).

The first layer of our model is employs same convolutions over the histograms of each timeseries. We apply a max-pooling operation over the *rows* of the resulting feature maps and set $\hat{c}_i = \max_k\{c_{k,i}\}$ as the final output feature of dimension i. This means we are pooling over histograms (but not vertically, *i.e.* across time), and that there is one output feature per timestep. Furthermore, we use two-dimensional filters that incorporate some information from past and future histograms as well as the present time step they are being pooled over. These convolutional features are passed to the next layer, an LSTM.

We employ an LSTM that takes convolutional features as input. We feed the complete timeseries of inputs into this component and then attach a fully connected layer to its final hidden state.

We use the final fully connected layer to project the LSTM's final hidden state vector into a probability distribution over the output space. Since we employ a binary labeling scheme (section 4.1), this distribution corresponds to the probability that an outbreak of rust occurred in some geographic region.

We evaluated all models with 20-fold nested cross-validation. We used the inner loops to cross-validate over model architectures and hyperparameters. Optimal performance was achieved with minibatches of 16, dropout at a rate of 0.5, 40 histogram buckets, 16 filters of size 3×3, 1 unidirectional LSTM layer with 512 hidden cells, and 64-unit fully connected layer (Figure. 4). For the nonconvex optimization of model parameters, we use Adam [22] with a learning rate of 0.0003, which converges in less than an hour on an NVIDIA Titan X GPU.

We test our models with the nested cross-validation procedure's outer loop and evaluate with AUC, the area under the ROC curve.

6.2. Comparative Study

To test the efficacy of our proposed technique, we compared our deep features to the following baselines:

1. An l2-regularized logistic regression classifier with a maximum likelihood criterion and widely adopted spectral features ("ML classifier", Table 1). Spectral-index based classifiers are almost unanimously recommended for wheat rust monitoring and thus serve to exemplify the efficacy of "traditional" techniques in this domain [39, 35, 1].
2. Another ML classifier trained on plain histograms. This ML histogram classifier makes explicit any advantages deep features have over the plain bins of histogrammed imagery.

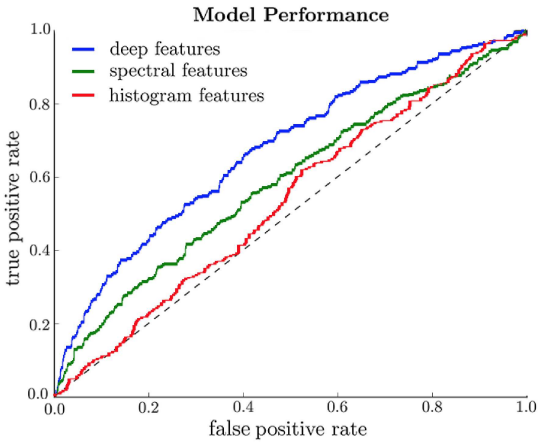

Figure 5. ROC curves for the proposed model (deep features), ML baseline with histogram features (raw histogram features), and ML baseline with spectral features (spectral features).

Results are shown in Figure 5. With an AUC of 0.67, our results demonstrate that the CNN and LSTM approach significantly outperforms that of traditional classifiers on raw histogram buckets (AUC of 0.53) and spectral indices (AUC of 0.58). Furthermore, our deep model maps these rust outbreaks with an accuracy of 76.53%. This degree of accuracy is competitive with state-of-the-art wheat disease mapping projects that leveraged intensive year-round field inspection (78%) [39] and high-resolution hand-held spectrometers on specially inoculated fields (82%) [35]. Note, though, the caveats to this comparison: accuracy is a less informative evaluation metric than AUC. Furthermore, we are generating region-level disease maps, whereas all prior work generates field-level predictions.

6.3. Predicting Outbreaks

Predicting future rust outbreaks is critical for food safety, public health, and pathogen mapping. To test the efficacy of our model in this setting, we trained and tested our the same models from section 5.1 on sub-sequences of the input $\boldsymbol{H}^{(1)}, \cdots, \boldsymbol{H}^{(r)}$ with $r < m$. In Ethiopia, wheat is planted in July and harvested by April of the following year. Therefore, we started with July's data and progressed through the season to predict outbreaks in an online manner.

Our results are presented in Figure 6. They suggest that learned features better distill the increasing complexity of temporal data. Early on in the season, all experimental variants perform poorly, with AUC's near 0.5. This is expected

Spectral feature	Definition	Formula
R_B	Blue reflectance (459 - 479 nm)	
R_G	Green reflectance (545 - 565 nm)	
R_R	Red reflectance (620 - 670 nm)	
R_{NIR}	Near-infrared reflectance (841 - 876 nm)	
GPP	Gross Primary Productivity [40]	
$NVDI$	Normalized Difference Vegetation Index [33]	$(R_{NIR} - R_R)/(R_{NIR} + R_R)$
$GNVDI$	Green Normalized Difference Vegetation Index [10]	$(R_{NIR} - R_G)/(R_{NIR} + R_G)$
TVI	Triangular vegetation index [3]	$0.5[120(R_{NIR} - R_G) - 200(R_R - R_G)]$
$SAVI$	Soil adjusted vegetation index [16]	$1.5(R_{NIR} - R_R)/(R_{NIR} + R_R + 1.5)$
$OSAVI$	Optimized soil adjusted vegetation index [31]	$(R_{NIR} - R_R)/(R_{NIR} + R_R + 0.16)$
$PSRI$	Plant senescence reflectance index [25]	$(R_R - R_B)/R_{NIR}$
MSR	Modified Simple Ratio [13]	$(R_{NIR}/R_R - 1)/(R_{NIR}/R_R + 1)^{0.5}$
NLI	Non-linear vegetation index [11]	$(R_{NIR}^2 - R_R)/(R_{NIR}^2 + R_R)$
$RDVI$	Re-normalized difference vegetation index [32]	$(R_{NIR} - R_R)/(R_{NIR} + R_R)^{0.5}$
SR	Simple ratio [2]	R_{NIR}/R_R
$CARI$	Clorophyll absorption ratio index [21]	$\lvert a \cdot 670 + R_R + b\rvert/\sqrt{a^2+1} \cdot (R_{NIR}/R_R)$, $a = (R_{NIR} - R_G)/150, b = R_G - a \cdot 550$
dX	First derivatives of all the above features with spectral change normalization [39]	$dX = (X_{later} - X_{earlier})/(X_{later} + X_{earlier})$

Table 1. Definitions of spectral features that were used in the baseline classifier. Each R variable correponds to some reflectance band, i.e. R_{NIR} for near-infarec, R_G for green, and so on.

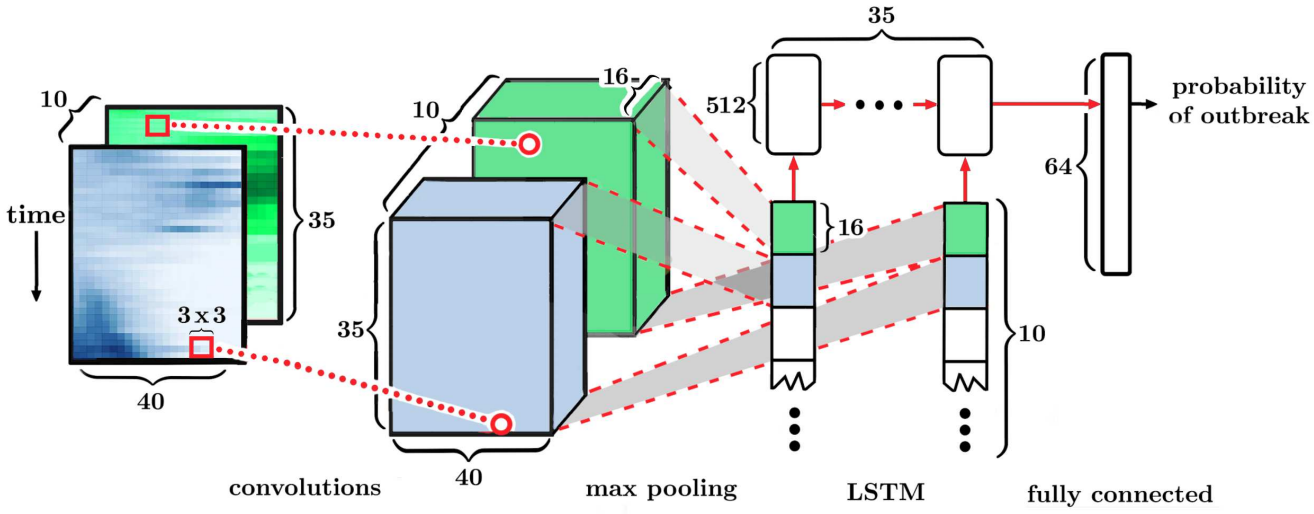

Figure 4. An illustration of our model architecture, explicitly showing all operations and dimensionalities. Matrices, vectors and tensors are depicted as boxes and cubes. Function applications are shown in red (dot product: "$\cdots$", pooling: "- - -", matrix multiplication: "$\rightarrow$"). The activations of the last fully connected layer are automatically learned features of the data. Our results suggest that within the domain of discovering Ethiopian rust outbreaks with satellite imagery, these deep features outperform the handcrafted spectral indices of Table 1.

as we are too close to the start of the season to make informed predictions. As we subsume time, all the models improve, and the gap between our deep learning model and baselines widens. Interestingly, this gap spikes near the end

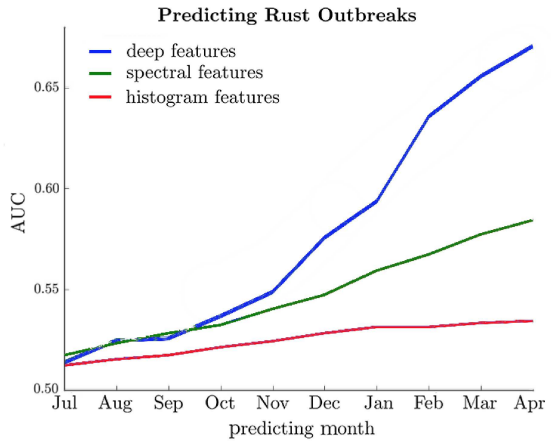

Figure 6. Classification performance in AUC as the survey data is extended year by year.

of the Ethiopian dry season (Jan - Mar). This may be due to nuanced differences between (a) dried-out regions hostile to fungus and (b) wetter, fungus-prone regions that the deep features were best able to capture.

6.4. Spectral Band Ablation

Most traditional features used in remote sensing applications rely on a subset of spectral bands, primarily band 1 (red, $620-670$nm) and 2 (near-infrared, $841-876$nm). It is possible that automatically learned features are capable of drawing informative structure from overlooked bands. To test this hypothesis, we excised each band from our dataset, trained on the remainder, and evaluated the loss of quality in the resulting features (Figure 7).

It is evident that deep features are capable of distilling information from every band. It is interesting to note that band 3 has stronger control over feature effectiveness than bands 1 or 2, which are more common in traditional spectral classifiers (Table 1). Furthermore, among surface reflectance bands, the oft-overlooked mid-infrared bands (5 and 6) have the strongest impact on feature quality. Day and night temperature (bands 8 and 9) appear to have similar impact. GPP (band 10) has the strongest control over feature quality, possibly due to to the dichotomy in photosynthetic signature between outbreak and non-outbreak regions (Figure 3).

6.5. Effects of Ongoing Study

Monitoring wheat rust necessitates the collection of ground truth data from actual wheat fields, the dearth of which is a major limiting factor of the proposed method's potency. Our data originate from a survey organized by the CIMMYT. The survey began in 2007 and surveyors visited an average of 88 Woredas per season. To better understand the potential of our method, we observe how performance

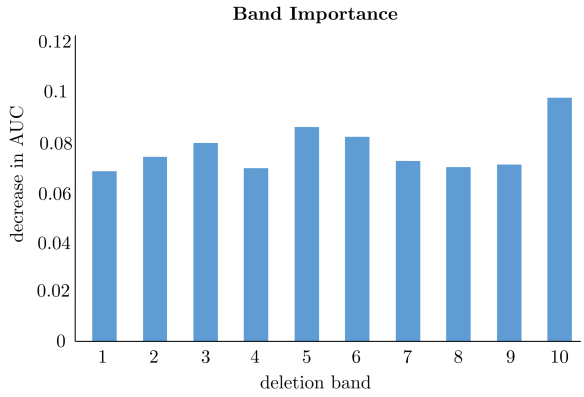

Figure 7. Decrease in classification performance as bands are dropped from consideration. Note that the more learned features are able to draw structure from a band, the more its removal will hurt performance. From left to right, bands are red (620 - 670 nm), near-infared (841 - 876 nm), blue (459 - 479 nm), green (545 - 565 nm), shallow infared (1230 - 1250 nm),- medium infared (1628 - 1652 nm), deep infared (2105 - 2155 nm), day temperature, night temperature, and gross photosynthetic product.

changes as we allow the model to train on more years of survey data (Figure 8).

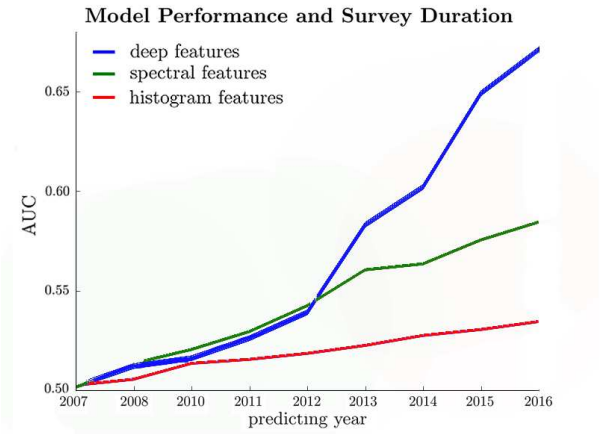

Figure 8. Classification performance in AUC as models are allowed to train on more months of data. Note the sudden increase in deep feature performance at the end of the Ethiopian dry season.

It is evident that more data boosts the performance of every technique. Our deep learning model, however, begins to separate itself from the rest after 2012. These results suggest that as surveys of this kind continue, all methods will be increasingly capable of identifying wheat rust, but that deep learning models may benefit the most.

7. Conclusion

We present a promising framework for predicting fungal outbreaks with hyperspectral satellite imagery and apply it to nine years of Ethiopian agricultural outcomes. This technique relies on unstructured surveys, which are cheaper and more scalable than the controlled experiments that dominate the space. The method is capable of real-time forecasting throughout the growing season. Our evidence suggests that its automatically learned features are more expressive than traditional spectral indices, and that this gap will widen in the coming years. It is our hope that this technique will provide the foundation for increased food security in the developing world.

The proposed technique is promising but imperfect. With an AUC of 0.67, its predictions are proibitively inaccurate for practical applicability. Future work might explore the incorporation of higher spatial- and temporal-resolution imagery, as well as transferring knowledge from models trained on related tasks like crop yield monitoring.

8. Acknowledgements

We thank Dave Hodson at CIMMYT for providing the field dataset and motivating the project. We gratefully acknowledge support from Stanfords Global Development and Poverty Initiative, and NSF grants 1651565 and 1522054 through subcontract 72954-10597.

References

[1] D. Ashourloo, M. R. Mobasheri, and A. Huete. Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. *Remote Sensing*, 6(6):5107–5123, 2014. 1, 2, 5

[2] F. Baret and G. Guyot. Potentials and limits of vegetation indices for lai and apar assessment. *Remote sensing of environment*, 35(2-3):161–173, 1991. 6

[3] N. H. Broge and E. Leblanc. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote sensing of environment*, 76(2):156–172, 2001. 6

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3

[5] R. Devadas, D. Lamb, S. Simpfendorfer, and D. Backhouse. Evaluating ten spectral vegetation indices for identifying rust infection in individual wheat leaves. *Precision Agriculture*, 10(6):459–470, 2009. 2

[6] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. 2

[7] R. FAO et al. Faostat database. *Food and Agriculture Organization of the United Nations, Rome, Italy*, 2013. 1

[8] D. Fink, T. Damoulas, and J. Dave. Adaptive spatio-temporal exploratory models: Hemisphere-wide species distributions from massively crowdsourced ebird data. In *AAAI*, 2013. 1

[9] S. Fritz, L. See, I. McCallum, L. You, A. Bun, E. Moltchanova, M. Duerauer, F. Albrecht, C. Schill, C. Perger, et al. Mapping global cropland and field size. *Global change biology*, 21(5):1980–1992, 2015. 3

[10] A. A. Gitelson and M. N. Merzlyak. Remote estimation of chlorophyll content in higher plant leaves. *International Journal of Remote Sensing*, 18(12):2691–2697, 1997. 6

[11] N. S. Goel and W. Qin. Influences of canopy architecture on relationships between various vegetation indices and lai and fpar: A computer simulation. *Remote Sensing Reviews*, 10(4):309–347, 1994. 6

[12] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 3

[13] D. Haboudane, J. R. Miller, E. Pattey, P. J. Zarco-Tejada, and I. B. Strachan. Hyperspectral vegetation indices and novel algorithms for predicting green lai of crop canopies: Modeling and validation in the context of precision agriculture. *Remote sensing of environment*, 90(3):337–352, 2004. 2, 6

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[15] W. Huang, D. W. Lamb, Z. Niu, Y. Zhang, L. Liu, and J. Wang. Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. *Precision Agriculture*, 8(4-5):187–197, 2007. 2

[16] A. R. Huete. A soil-adjusted vegetation index (savi). *Remote sensing of environment*, 25(3):295–309, 1988. 6

[17] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 1

[18] C. Justice, J. Townshend, E. Vermote, E. Masuoka, R. Wolfe, N. Saleous, D. Roy, and J. Morisette. An overview of modis land data processing and product status. *Remote sensing of Environment*, 83(1):3–15, 2002. 3

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[20] S. Kelling, J. Gerbracht, D. Fink, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, and C. P. Gomes. ebird: A human/computer learning network for biodiversity conservation and research. In *IAAI*. Citeseer, 2012. 1

[21] M. S. Kim, C. Daughtry, E. Chappelle, J. McMurtrey, and C. Walthall. The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (a par). 1994. 6

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[24] L. Liu, X. Song, C. Li, L. Qi, W. Huang, and J. Wang. Monitoring and evaluation of the diseases of and yield winter wheat from multi-temporal remotely-sensed data. *Trans-*

actions of the Chinese Society of Agricultural Engineering, 25(1):137–143, 2009. 2

[25] M. N. Merzlyak, A. A. Gitelson, O. B. Chivkunova, and V. Y. Rakitin. Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia plantarum*, 106(1):135–141, 1999. 6

[26] D. Moshou, C. Bravo, J. West, S. Wahlen, A. McCartney, and H. Ramon. Automatic detection of yellow rustin wheat using reflectance measurements and neural networks. *Computers and electronics in agriculture*, 44(3):173–188, 2004. 2

[27] P. Olivera, M. Newcomb, L. J. Szabo, M. Rouse, J. Johnson, S. Gale, D. G. Luster, D. Hodson, J. A. Cox, L. Burgin, et al. Phenotypic and genotypic characterization of race tkttf of puccinia graminis f. sp. tritici that caused a wheat stem rust epidemic in southern ethiopia in 2013–14. *Phytopathology*, 105(7):917–928, 2015. 1

[28] H. Qiao, B. Xia, X. Ma, D. Cheng, and Y. Zhou. Identification of damage by diseases and insect pests in winter wheat. *Journal of Triticeae Crops*, 4:041, 2010. 2

[29] N. Quarmby, M. Milnes, T. Hindle, and N. Silleos. The use of multi-temporal ndvi measurements from avhrr data for crop yield estimation and prediction. *International Journal of Remote Sensing*, 14(2):199–210, 1993. 2

[30] C. Robert, M.-O. Bancal, B. Ney, and C. Lannou. Wheat leaf photosynthesis loss due to leaf rust, with respect to lesion development and leaf nitrogen status. *New Phytologist*, 165(1):227–241, 2005. 1

[31] G. Rondeaux, M. Steven, and F. Baret. Optimization of soil-adjusted vegetation indices. *Remote sensing of environment*, 55(2):95–107, 1996. 6

[32] J.-L. Roujean and F.-M. Breon. Estimating par absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment*, 51(3):375–384, 1995. 6

[33] J. Rouse Jr, R. Haas, J. Schell, and D. Deering. Monitoring vegetation systems in the great plains with erts. 1974. 6

[34] R. P. Singh, D. P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S. A. Herrera-Foessel, and R. W. Ward. Will stem rust destroy the world's wheat crop? *Advances in agronomy*, 98:271–309, 2008. 1

[35] H. Wang, F. Qin, Q. Liu, L. Ruan, R. Wang, Z. Ma, X. Li, P. Cheng, and H. Wang. Identification and disease index inversion of wheat stripe rust and wheat leaf rust based on hyperspectral data at canopy level. *Journal of Spectroscopy*, 2015, 2015. 1, 2, 5

[36] H. Wenjiang, H. Muyi, L. Liangyun, W. Jihua, Z. Chunjiang, and W. Jindi. Inversion of the severity of winter wheat yellow rust using proper hyper spectral index. *Transactions of the Chinese Society of Agricultural Engineering*, 21(4):97–103, 2005. 2

[37] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098*, 2015. 1

[38] J. You, X. Li, M. Low, D. Lobell, and S. Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. 2017. 4

[39] J. Zhang, R. Pu, L. Yuan, J. Wang, W. Huang, and G. Yang. Monitoring powdery mildew of winter wheat by using moderate resolution multi-temporal satellite imagery. *PloS one*, 9(4):e93107, 2014. 1, 2, 5, 6

[40] M. Zhao, F. A. Heinsch, R. R. Nemani, and S. W. Running. Improvements of the modis terrestrial gross and net primary production global data set. *Remote sensing of Environment*, 95(2):164–176, 2005. 6