# Prediction of *Sorghum bicolor* Genotype from In-situ Images Using Autoencoder-identified SNPs

Mihael Cudic
*Dept. of Engineering Science*
*University of Oxford*
*Oxford, United Kingdom*
*mihael.cudic@balliol.ox.ac.uk*

Harjatin S. Baweja, Tanvir Parhar, and Stephen T. Nuske
*Robotics Institute*
*Carnegie Mellon University*
*Pittsburgh, United States*
*harjatis@andrew.cmu.edu, ptanvir@andrew.cmu.edu, snuske@andrew.cmu.edu*

***Abstract*—Extensive genetic and phenotypic research is necessary for any effective plant breeding program. Such studies, however, require an immense amount of time and resources. In order to expedite the breeding process, we provide a novel method for rapid genotype prediction using in-situ images of plants. In this method, significant single nucleotide polymorphisms (SNPs) are first identified using a novel autoencoder framework with the goal of being more robust to false positive associations than standard genome wide association studies (GWAS). On-field images of various plant varieties are then used to train Convolutional Neural Networks (CNNs) to predict candidate alleles and validate phenotypic relationships. This image-based system allows for easy use on new plant varieties to gain real-time genetic information for better harvest prediction. The feasibility of our method for rapid genotype prediction was demonstrated on 345 *Sorghum bicolor* varieties with corresponding uncontrolled images 60 days after seed planting. Our autoencoder identified 4 significant SNPs that had an average allele classification accuracy of 70.58% on 68 previously unseen plant varieties.**

***Keywords*-Genotype Prediction, Genome Wide Association Studies, Locally Connected Autoencoder, Convolutional Neural Networks**

## I. Introduction

The genome encodes all information necessary to fully reconstruct the biological properties of an organism. Alcohol dehydrogenase in maize (*Zea mays L.*), for example, is caused by the single-allele *ADHI* gene [1]. *D8* and *Mpl1* mutations in maize have likewise been linked to dwarfism traits that drastically reduce yield [2]. By isolating these significant alleles, further research can be done to target specific locations and augment genomes to potentially alter certain properties in crops.

Plant breeding programs therefore rely on understanding a crop's genotype to better inform breeders which plant varieties to cross. This process known as genomic selection (GS) results in cultivars with unprecedented trait expressions, vastly improving upon simpler breeding approaches such as mutant selection [3]. However, mapping the genome of a new organism requires years of manual, computational, and biological effort and even more time is required to fully optimize a crop. To expedite GS breeding programs, we introduce a methodology for rapid prediction of alleles thought to be highly associated to phenotypes [4]. Through this method, we demonstrate its genetic predictive power on new plant varieties and eliminate the need for continuous extensive genetic sequencing. This image-based system is a light-weight easy-to-use system that can additionally be used by everyday farmers lacking the resources of well-established breeding institutions.

Rapid genotype prediction was accomplished by first identifying significant single nucleotide polymorphisms (SNPs) through a novel genome wide association study (GWAS) using an autoencoder. To the best of our knowledge, we are the first to demonstrate the usefulness of an autoencoder in identifying significant SNPs within a small-sample high-dimensional allele panel containing sub-populations. Unlike traditional GWAS algorithms, an autoencoder is trained unsupervised and learns features in the allele space negating any potential for high false positive association rates [5]. We show that some of the extracted features in the encoding layer separate phenotypes into statistically significant clusters [6]. The relative contribution, and thus the significance, of each SNP to phenotype separability was then calculated by multiplying backpropagated error gradients from discriminating encoding dimensions with the SNP's minor allele frequency.

SNPs scoring highest on this metric were identified and their significance was further validated by mapping images of in-situ plants to their respective genotype using convolutional neural networks (CNNs). Over the years, CNNs have proved to be highly successful in establishing benchmarks in several image analysis domains ranging from classification [7], segmentation [8] and detection [9]. By corroborating correlations between on-field image data of crops and identified SNPs, more confidence regarding the alleles' significance is gained. Additionally, activation maps from the CNNs provide useful analysis to determine the expression of the SNP in question [10]. The results of the analysis described herein can be easily distributed to the public for rapid in-situ allele prediction circumventing the need for large scale DNA sequencing.

Our study was performed on *Sorghum bicolor* accessions provided by the Institute for Translational Genetics at Clemson University, South Carolina [11] as well as field images corresponding to each accession collected by our group. Sorghum has emerged as a leading candidate for bioenergy feedstock due to its ease of modification, diversity, and efficiency. If more knowledge is known about Sorghum, different varieties can be crossbred or genetically modified to maximize biofuel production. The allele panels contain 345 accessions with 232,303 SNPs for each accession and 2,648 field images were taken over all sorghum varieties 60 days after planting. Using our autoencoder, we have shown that we can isolate SNPs directly corresponding to various Sorghum phenotypes including fresh plant weight, stalk height, and lignin percentage. Four new markers - S1_18096891, S2_2633689, S2_2793224, and S6_7123769 - were identified and achieved an average allele classification accuracy of 70.58% using images of 68 previously unseen Sorghum varieties.

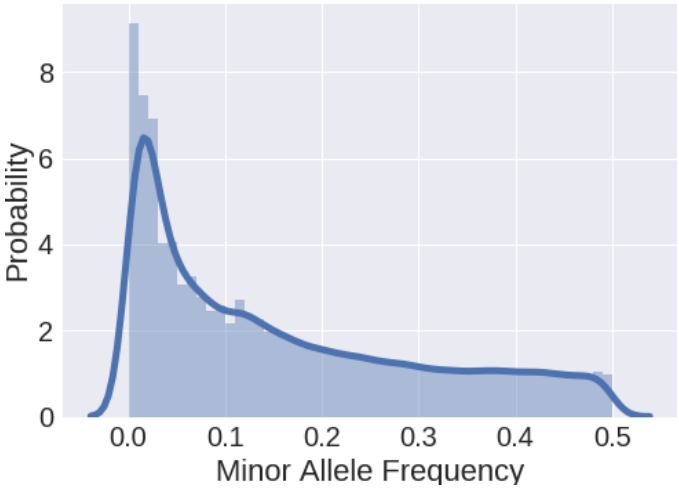



Figure 1: The distribution of minor allele frequencies across all 232,303 SNPs. SNPs with a minor allele frequency below 0.1 were filtered out of our analysis to guarantee more evenly distributed alleles.

## II. Related Works

### A. Genome Wide Association Studies

The simplest method to perform GWAS is through case-controlled studies that use statistical tests for independence by analyzing each individual SNP's contribution to a phenotype. However, treating each SNP as independent leads to high false positive rates as it ignores population structure and genetic relatedness. Structured association models were proposed to account for inbreed populations by requiring similar statistical tests to be consistent throughout each subpopulation [12]. Likewise, principal component analysis (PCA) can correct for biases from predefined structure [13].

These solutions only partially account for genetic similarities and, therefore, more sophisticated models are necessary. By incorporating a kinship matrix describing pairwise genetic relatedness between all similar individuals, Mixed Linear Models have been shown to be successful in associating quantitative phenotypes to genotypes given population structure [14]. More popular GWAS algorithms include the Efficient Mixed-Model Association (EMMA) algorithm and its derivative algorithms due to their computational efficiency and ability to account for internal structures [5][15][16]. Compressed Mixed Linear Models (CMLM) are also used to analyze high-dimensional allele panels when further reductions in the algorithm's time-complexity is required [17].

More recently, deep learning has been increasingly applied to bioinformatics because of its notable success in various fields such as computer vision and game AIs [18] [19]. DeepSea, for example, uses convolutions on a DNA sequence of noncoding genomic variants to predict traits with single nucleotide sensitivity [20]. Similarly, deepWAS, an iteration of DeepSea, performs GWAS with the aim of associating SNPs to major depressive disorders [21]. DNA sequence patterns that correlate to phenotype classes can additionally be analyzed using convolutions as shown by Lanchantin et al. [22]. These methods, however, are heavily prone to overfitting and biases from subpopulation specimen.

### B. Unsupervised Learning for Genetics

Alternatively, unsupervised learning approaches can learn the internal structure of the data and extract features from complex systems as well. Le et al. demonstrated that a deep autoencoder is able to separate features, such as a human or a cat, using 10 million images with no labels [23]. A similar application of autoencoders has also been implemented in genetics. Tan et al. utilized Denoising Autoencoders (DAs) to classify breast cancer tumors by extracting relevant features in the encoding layer [24]. Likewise, ADAGE shows a methodology for using autoencoders on gene expression data and demonstrates the ability to isolate high contributing inputs for further analysis [25]. Although unsupervised methods have primarily analyzed gene expression data, Montañez et al. were the first to show that autoencoders can extract relevant features from SNPs [26]. In this study, stacked autoencoders reduced the dimensionality of allele panels containing 2,465 SNPs thought to be related to obesity. Then, a feed forward artificial neural network was trained to predict extreme obesity from the encoded panels. Even with just 50 encoding nodes, Montañez et al. found that there was sufficient information present for accurate obesity classification. With this in mind, we look to extend this methodology to incorporate larger allele panels and directly identify which SNPs contribute to genotype.

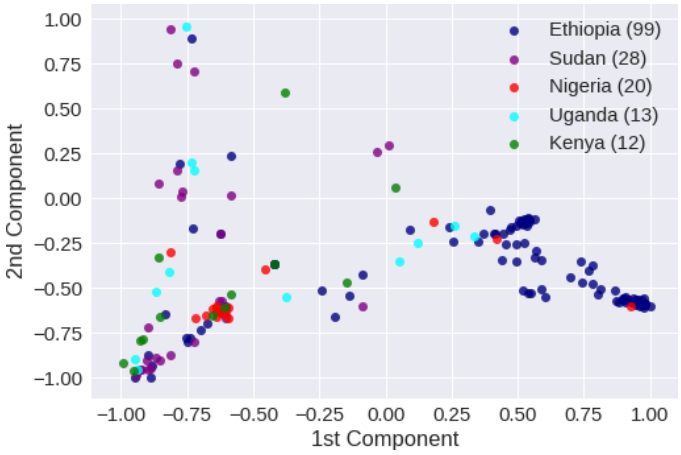
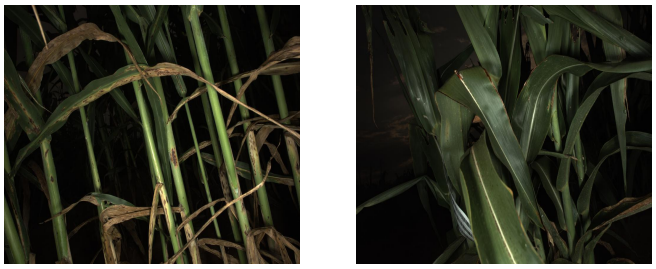



Figure 2: The first and second normalized principal components for cultivars originating from the 5 most reoccurring regions: Ethiopia, Sudan, Nigeria, Uganda, Kenya. PCA was performed over 125,980 SNPs for all 345 cultivars. The number of cultivars from each region can be seen in the legend.



(a) (b)

Figure 3: Image (a) shows an example of an ideal image for allele classification task and (b) shows an occluded image devoid of much useful information

## III. Datasets

The *Sorghum bicolor* data collected consists of 345 cultivar accessions with 232,303 SNPs across 10 chromosomes. Although the vast majority of sampled SNPs have no known associations to phenotype, some SNPs come from previously mapped dwarfism genes (*DW1*) and maturity genes (*MA1*, *MA3*, and *MA6*) [11]. All missing SNP values were imputed resulting in data falling into three classes: major allele, mixed, and minor allele. 95.62% of the data can be described either in the first or last class. However, the distribution of major and minor alleles is heavily skewed as seen in figure 1. Because it is hard to create generalizable associations given a few instances of an allele, we only used SNPs with a minor allele frequency greater than 0.1 leaving 125,980 SNPs for analysis.

In addition, the Sorghum originated from 37 locations all over the world with only 87 cultivars originating from unknown locations. Roughly half of the Sorghum varieties originated from only 5 locations: Ethiopia, Sudan, Nigeria, Uganda, and Kenya. Because the Sorghum was not sampled evenly throughout the world, there lies a high chance of evolutionary inbreeding and thus genetic relatedness. Subpopulations were partially visualized by plotting the first and second principal components from cultivars originating from these 5 regions. As shown in figure 2, Ethiopian and Nigerian plant varieties cluster while only some Sudanese plant varieties cluster. However, not all subpopulations can be captured using PCA as PCA performs linear dimensionality reduction and principal components bias toward subpopulations that occur the most.

Corresponding images for each 345 Sorghum accessions were also taken as part of our groups plant phenotyping work at Carnegie Mellon University [27]. Images were taken 60 days after planting in July 2016 on a Sorghum research field in Pendleton, South Carolina. A Robot Operating System, 9MP stereo-camera pair with 8mm focal length, and a high powered flash triggered at 3Hz were used to capture image data. The sensor was driven at approximately 0.05m/s on our Robotics platform [28] maintaining a distance of approximately 0.8m from the plant growth. Thus, we acquired roughly 10 images per sorghum variety resulting in 2,648 images in total. As shown in figure 3, these images were taken in an uncontrolled setting with some images having high stalk occlusion. It is therefore not guaranteed that every image contains all the information necessary for genotype prediction.

## IV. Methodology

Although we were able to reduce the dimensionality of the allele panel from 232,303 SNPs to 125,980 SNPs, the input dimensionality was still significantly larger than the number of samples. This small-sample high-dimensionality combination limits any significant analysis as it inflates the false positive rate and the possibility for overfitting when using conventional GWAS algorithms. Thus, we wanted to use an unsupervised approach to reduce the dimensionality by extracting features in the panel.

Autoencoders provide non-linear dimensionality reduction and have been successful in many instances [23]. Because our 125,980 input dimension was too large to construct a fully connected autoencoder, we used localized connections in the first and final layers of the autoencoder as shown in figure 4. Similar architectures have been developed before, but none have taken advantage of locally connected layers' ability to downsample and upsample with precision [23][29]. Unlike regular convolutional autoencoders that look to encode data with spatial correlation, each dimension of the allele panel must be treated as independent. This biologically inspired structure bodes well for genetic data as adjacent markers are more likely to be correlated with one another than well separated markers. Thus, each locally connected

node can hopefully represent the associated SNPs more accurately. Overlap between locally connected nodes allow for more flexibility and information, providing a medium between reduced parameters and fully connected layers. Overlap is especially important in the upsampling layer as it allows output nodes to form more than one connection. Finally, the middle two layers of the system are fully connected to ensure the allele panels are encoding global information.

To prevent the locally connected autoencoder (LCA) from learning a one-to-one mapping, the encoding layer must be less than 345 nodes. We did not know how many encoding nodes would generate the best separation among phenotypes, so we experimented with 256, 128, and 64 encoding nodes. Sigmoid and tanh activation functions were tested separately in the encoding layer to create discrete clusters for further analysis. In addition, we used leaky ReLU functions with an alpha of 0.05 throughout the LCA to ensure that no nodes were lost. A sigmoid activation function was used for the output allele classification.

Inputs and outputs were encoded as a 0, 0.5, and 1 for major allele, mixed, and minor allele respectively. The training set consisted of all 345 cultivars and the LCA was trained with binary crossentropy for all 125,980 SNPs. Mixed outputs were not used to calculate classification scores and their errors were zeroed during backpropagation as mixed outputs were not binary. Additionally, a small step size of 5.0e-5, a decay rate of 0.98 for every 1,500 updates, Adam optimization, and a batch size of 8 were used to prevent large errors accumulating from our high-dimensional output and smoothen the learning curve [30]. Training was stopped after 30 epochs of no improvement in allele decoding classification on the training set.

After training, encoding dimensions were then binarized and saved for every cultivar. Each encoding node was tested across all phenotypes to determine their ability to cluster into significantly different distributions. Because we were interested in mean-based separation, we performed p-tests between clustered distributions to quantify the significance of the separation. All clustered distributions with p-tests below 5.0e-5 were examined manually as an additional test. Once the correlation between the encoding node and phenotype was confirmed, contribution of each allele was calculated by backpropagating a 1.0 error for the node in question as outline by Dimopoulos et al. [6]. Due to the skew in minor allele frequency, each calculated sensitivity was also multiplied by the minor allele frequency to generate an average allele contribution across all Sorghum varieties.

For the next step, we trained CNNs to predict the autoencoder-identified SNPs of each cultivar from corresponding image data. The CNNs had 5 convolutional layers with 3x3 kernels and each convolutional layer was followed by ReLU nonlinearities, a pooling layer, and Local Response Normalization (LRN). After the 5 convolutional layers, two fully connected layers of 4,096 neurons and a final Softmax layer for a two-way classification followed. We used Adam optimization with a base learning rate of 1e-04 and a decay rate of 0.9 for every 10,000 updates [30]. The images were split 80% to 20% for training and validation with separate gene accessions. Thus, validation images were of plant varieties not seen in training which ensures that the CNN is able to generalize to new plant varieties.

Activation maps from the trained CNNs also provided better insight into the learning process and helped determine possible SNP expression [10]. Only the activation maps of first few convolutional layers were used as they are known to learn simpler and more interpretable features. Additionally, the receptive fields for the first few convolutional layers are much more resolved for each neuron than those in the later layers. This allows us to better determine which parts of the image are associated to the SNP.

## V. Results

We were able to identify 5 significant SNPs for genotype prediction with our methodology. Unlike conventional GWAS algorithms, the LCA looks to understand the global structure of the DNA sequence and we can be more confident in the associations found for all 5 SNPs. After analyzing the correlations between in-situ Sorghum images and alleles, we were able to predict genotype on 4 SNPs with 70.58% accuracy on previously unseen Sorghum varieties.

Table I: Performance of LCA with various encoding hyperparameters

| Encoding Nodes | Accuracy sigmoid | Accuracy tanh | p-test sigmoid | p-test tanh |
|---|---|---|---|---|
| 64 | 84.05% | 83.73% | 4.10e-18 | 6.10e-17 |
| 128 | 87.56% | 85.05% | 6.88e-14 | 1.36e-17 |
| 256 | 89.48% | 85.61% | 1.03e-15 | 1.73e-17 |

Several activation functions and sizes were tested to determine the optimal hyperparameters for the encoding layer. As shown in table II, more encoding nodes led to better reconstruction classification as each node encoded more resolved features. Sigmoid activations additionally received a maximum reconstruction accuracy of 89.48%, performing better than tanh activations. When analyzing activations across all encoding nodes, sigmoid activations seemed to be more equally distributed than their tanh counterpart and thus better optimized the use of each node.

Our ultimate goal, however, was not to encode allele panels with high accuracy but rather to extract meaningful features that can be expressed in the phenotype space. To quantify the significance of the features learned, p-tests were performed on all cluster separations. The average p-test score over the 10 best separations for each LCA is shown in table II as well. Unlike reconstruction classification, we

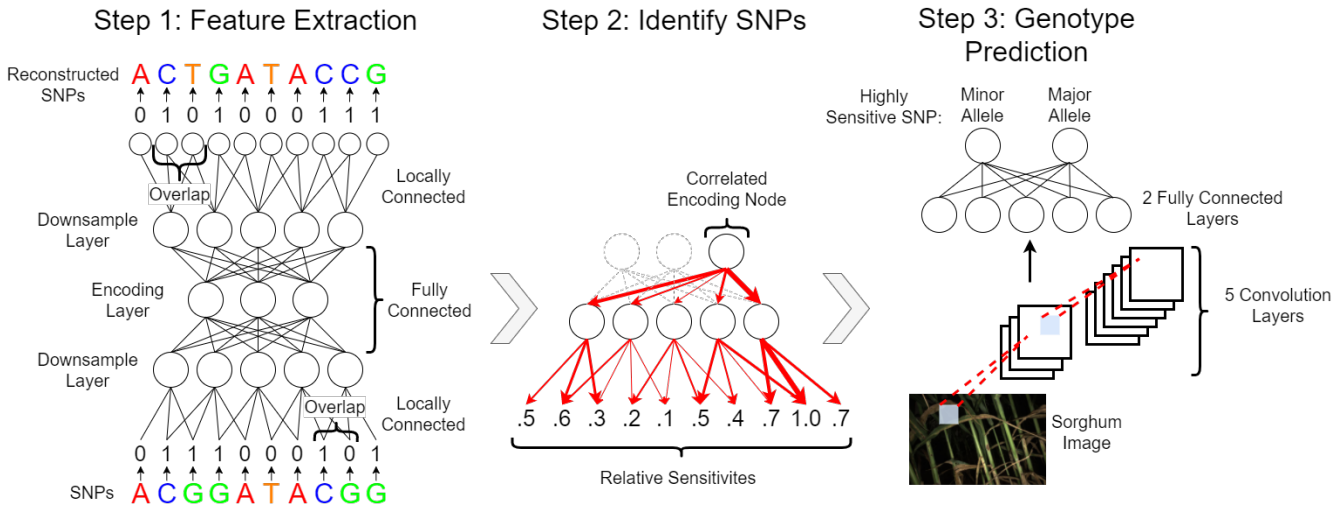



Figure 4: Summarized methodology for rapid genotype prediction. 1) A locally connected autoencoder was used to extract features from the allele panel. Only the first and last layers are locally connected to function as downsampling and upsampling layers respectively. Each downsample node corresponds to 40 core SNPs but has an additional overlap of 100 SNPs to the left and right of the core grouping resulting in a 3,150 sized layer. The upsample layer follows an inverse structure. The encoding layer was tested using various dimension sizes - 64, 128, 256 - and activations - sigmoid or tanh. The outputs had sigmoid activation functions. 2) Significant SNPs were identified by backpropagating the error gradient from correlated encoding nodes. This value was then multiplied by the minor allele frequency and all values were normalized to assign each SNP a relative contribution to phenotype separability. 3) Significant SNPs were isolated and on-field images were used to predict the allele with a CNN. The CNN contains 5 convolutional layers and 2 fully connected layers with ReLU activations. The output is a softmax layer containing two nodes for minor and major allele

can see that tanh activations generally perform best. This is likely due to the skewed activations learned as equal subpopulations are rare to find in nature. Because all p-test scores were very low, the best 50 separations were visually analyzed to confirm phenotype separation. 256 tanh nodes were identified to best separate phenotype data and were used for the remainder of this study.

The most significant phenotype separations found can be seen in figure 5. Encoding dimensions 82 and 90 separated phenotypes measuring total fresh plant weight, stalk height, and lignin percentage while encoding dimensions 30 seemed to correlate well with anthesis date and NFC percentage. Chi-squared tests were performed based of the cultivars origin to ensure that the clusters formed were independent of subpopulation structure. All 4 separations had statistically insignificant p-values showing that these features are true to the trait and not the region, demonstrating the LCA's ability to take into account unknown population structures.

The LCA was reinitialized and trained several times with each experiment resulting in the same phenotype separations. This ensures that the LCA is extracting real features and not finding random separations. The false positive rate of our system was also gauged by analyzing a new randomized allele panel with the same dimensions and minor allele frequencies as the *Sorghum bicolor* panel. The LCA was only able to learn the baseline reconstruction accuracy of 73.4% and thus found no phenotype separation or significant SNP. Contrastingly, a CMLM analysis on the same randomized allele panel identified roughly 0.967% of SNPs to be significantly associated with fresh weight when using a 0.01 p-value cutoff. These results are of no surprise as the LCA cannot encode random data and relies on multiple corresponding alleles to extract features. CMLMs, on the other hand, only look at a single SNP's correlation to a quantitative phenotype and will always find associations in high-dimensional allele panels. As a result, this validates the significance of the phenotype separations found as the LCA is able to better avoid false positives.

All phenotypes analyzed are of interest for further analysis, but for the sake of this study we specifically looked to encoding nodes 82 and 90 as they relate to biofuel production. By backpropagating the gradient to each SNP and multiplying it by the minor allele frequency, we can perform a sensitivity analysis for each encoding node. Unlike CMLM where SNPs are given a p-value describing its significance, we assign relative normalized sensitivities to each SNP as as shown in figure 6. There is no intuitive cutoff for separating significant and non-significant SNPs; however, when looking specifically at these two nodes, only 5 SNPs had relative sensitivities greater than 0.95: S1_18096891,

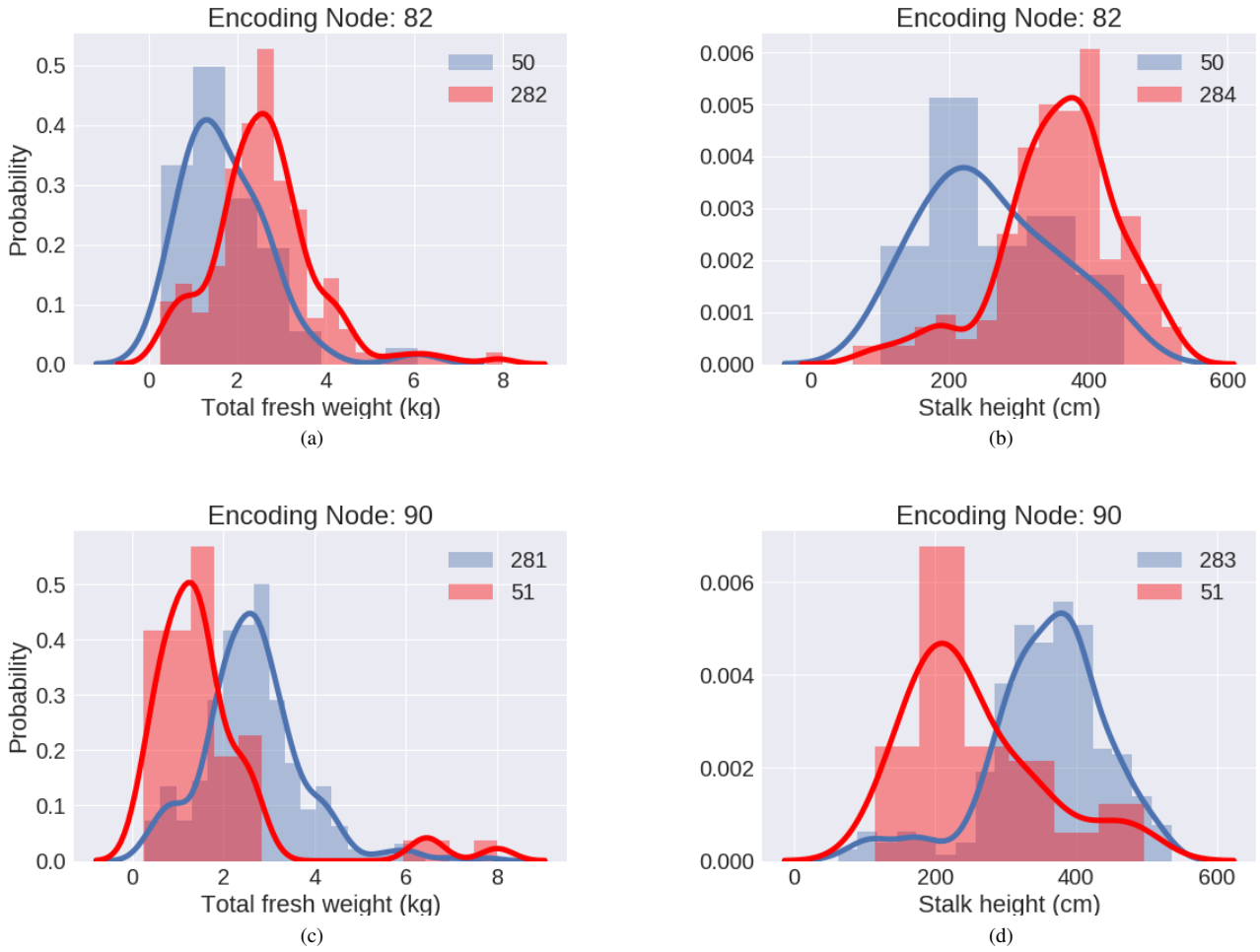



Figure 5: Clusters created by encoding node 82 and 90 on total fresh plant weight (a and c) and stalk height (b and d) data. The cultivars per distribution can be seen in the legend. The phenotype separation was achieved through an unsupervised autoencoder demonstrating the systems ability to extract meaningful features.

S2_2633689, S2_2793224, S6_61098274, and S6_7123769. A cutoff of 0.95 was thus implemented for this study even though SNPs slightly below might still be of interest. Only 4 locations were further analyzed due to the a high minor allele frequency in the S6_61098274 SNP.

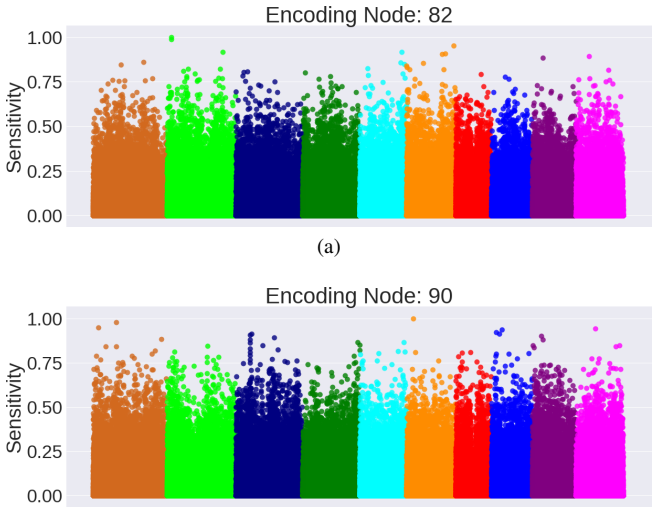
The 2,648 sorghum images of 277 Sorghum varieties were then used to train a CNN to predict these 4 SNPs. We can see the classification accuracy for each individual image on 68 new Sorghum varieties in table II. Since all images of a particular variety correspond to the same gene accession and since some images do not contain enough information due to occlusions, we took 5 images from a sorghum variety at random for an aggregate allele classification. This increases the likelihood that some images will contain the required information to more accurately assign a genotype. When taking the consensus classification from 5 images of the same plant type, we see an increase in allele classification accuracy from 65.9% to 70.6%.

To provide context on this performance, we attempted to predict alleles from 5 random SNP locations with minor allele frequencies above 0.45. The CNN achieved an average allele accuracy of just 51.17% and this result served as a lower-bound baseline. The experiment was repeated with an SNP of known association to the maturity *MA3* gene to serve as a reference for the maximum possible genotype classification on our image dataset. This SNP located at S1_60845833 has a 0.416 minor allele frequency and achieved a classification accuracy of 78.2%. The performance is only slightly better than for our autoencoder-identified SNPs; however, this can be partially explained by the greater major allele frequency in the S1_60845833 SNP. Thus, because the newly identified SNPs approach the 78.2% ceiling and perform much better than randomly selected alleles, we can be more confident in their association and

Table II: Accuracy of Convolutional Neural Network for 4 SNP locations with baseline split between major and minor allele

| SNP Location | Accuracy with one image | Accuracy with five images | Minor allele frequency |
|---|---|---|---|
| 5 Random SNPs* | 51.17% | 51.3% | 0.50 |
| S1_60845833** | 72.3% | 78.2% | 0.41 |
| S1_18096891 | 66.4% | 72.4% | 0.46 |
| S2_2633689 | 62.2% | 66.8% | 0.49 |
| S2_2793224 | 66.8% | 71.5% | 0.50 |
| S6_7123769 | 64.3% | 71.6% | 0.50 |

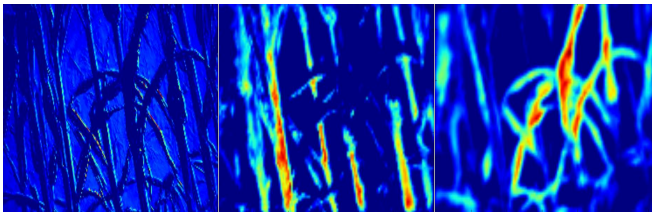
*5 random SNP locations were chosen with a minor allele frequency above 0.45. The average values are shown.
**S1_60845833 is an allele location with known association and thus serves to form our performance upper-bound.




(a)

(b)

Figure 6: Manhattan plots of the SNPs relative sensitivity to encoding node 82 and 90. Alleles with sensitivities above 0.95 were isolated as significant SNPs for further analysis. Although both plots highlight different regions, we can see a few SNPs have significantly higher sensitives than the rest: S1_18096891, S2_2633689, S2_2793224, S6_61098274, and S6_7123769. Note: each color corresponds to a new chromosome.



(a) (b) (c)

Figure 7: Activation maps of images show that certain learned features do interpretable image segmentation. For instance, figure (a) shows how the image's background has been segmented by learned features, figure (b) shows an accurate segmentation of sorghum stalks, and figure (c) shows a segmented image of plant leaves.

can use our trained CNNs for on-field genotype prediction. Other GWAS methods were excluded from experimental comparisons since they are all supervised and will therefore find SNPs that separate phenotypes well regardless of their biological relevance.

To further understand what the network is learning, we also visualized the activation maps at different convolutional layers in our network. As shown in figure 7, we can see interpretable features learned throughout the CNN. Certain filters in the first few convolutional layers learn to segment the image into background, stalks, and leaves . Because the SNPs identified correlate to total fresh weight, it makes sense for the network to differentiate between leaves and stalk as both differ in density.

## VI. Conclusion

In this study, we propose a novel methodology to predict genotype on new plant varieties using in-situ images. These systems can then be implemented by all institutions for cost-effective and rapid genotype prediction to expedite breeding processes. This was achieved by first identifying and validating genotype associations using LCAs and CNNs respectively. We show that LCAs are able to extract meaningful features in the allele panel that separate in the phenotype space. Unlike conventional GWAS algorithms, this method avoids any possibility for overfitting and high false positive rates while still being able to account for subpopulation structures. Then, the significance of SNPs are validated using CNNs through allele classification and further analysis of activation maps. Using *Sorghum bicolor* allele panels and image data, we found 4 significant SNPs that correspond to fresh plant weight and stalk height and achieve an average genotype prediction of 70.58%. Although we only looked at 4 SNPs, there is reason to believe that many more SNPs with similar relative sensitives can be provide genotype prediction as well.

These results serve as a proof of concept for our methodology but further improvements can be made. Any GWAS algorithm has trouble analyzing high-dimensional low-sample accessions as it is much easier to find statistical anomalies. Although our LCA is better able to handle such data, having more samples would allow for more resolved features with potentially more pronounced phenotype separations. Additionally, our CNN was only trained on roughly 2,100 images. By capturing more images from different plant varieties, the CNN would be less likely to overfit as it would be forced to learn more generalizable features. Most importantly, the im-

ages taken are all from on-field and uncontrolled conditions. Some images have high occlusion resulting in a portion of the images not containing the required information for genotype classification. In the future, images should be taken in more controlled settings that incorporate the full context of the plant for increased accuracy in genotype prediction.

Handling SNPs with highly skewed minor allele frequencies is one of the major challenges of this technique. Because such SNPs occur frequently, we filtered nearly half of the SNPs in the original panel and lost potential candidate genes for classification. An algorithm averse to these skewed distributions must be considered for further research.

## ACKNOWLEDGMENT

Authors would like to thank the colleagues from Kresovich lab at Clemson University for their plant genetics/phenotype data resources and invaluable insights into this domain. Further thanks to the team at Carnegie Mellon University; George Kantor, Tim Mueller-Sim, Omeed Mirbod, Anjana Nellithimaru, Merrit Jenkins, Jay Patravalli and Justin Abel for their efforts assisting in collecting the imaging data in field. Joe Cornelius and ARAPA-e for their vision and financial support of this research under ARPA-e Award No. 1830-219-2020937.

## REFERENCES

[1] D. Schwartz, "The genetic control of alcohol dehydrogenase in maize: gene duplication and repression," *Proceedings of the National Academy of Sciences*, vol. 56, no. 5, pp. 1431–1436, 1966.

[2] N. P. Harberd and M. Freeling, "Genetics of dominant gibberellin-insensitive dwarfism in maize." *Genetics*, vol. 121, no. 4, pp. 827–838, 1989.

[3] M. Goddard and B. Hayes, "Genomic selection," *Journal of Animal breeding and Genetics*, vol. 124, no. 6, pp. 323–330, 2007.

[4] M. F. Minamikawa, K. Nonaka, E. Kaminuma, H. Kajiya-Kanegae, A. Onogi, S. Goto, T. Yoshioka, A. Imai, H. Hamada, T. Hayashi *et al.*, "Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits," *Scientific Reports*, vol. 7, 2017.

[5] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, no. 3, pp. 1709–1723, 2008.

[6] I. Dimopoulos, J. Chronopoulos, A. Chronopoulou-Sereli, and S. Lek, "Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in athens city (greece)," *Ecological modelling*, vol. 120, no. 2, pp. 157–165, 1999.

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

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[11] Z. W. Brenton, E. A. Cooper, M. T. Myers, R. E. Boyles, N. Shakoor, K. J. Zielinski, B. L. Rauh, W. C. Bridges, G. P. Morris, and S. Kresovich, "A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy," *Genetics*, vol. 204, no. 1, pp. 21–33, 2016.

[12] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly, "Association mapping in structured populations," *The American Journal of Human Genetics*, vol. 67, no. 1, pp. 170–181, 2000.

[13] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, p. 904, 2006.

[14] M. Malosetti, C. G. van der Linden, B. Vosman, and F. A. van Eeuwijk, "A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato," *Genetics*, vol. 175, no. 2, pp. 879–889, 2007.

[15] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nature genetics*, vol. 44, no. 7, pp. 821–824, 2012.

[16] H. M. Kang, J. H. Sul, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, E. Eskin *et al.*, "Variance component model to account for sample structure in genome-wide association studies," *Nature genetics*, vol. 42, no. 4, pp. 348–354, 2010.

[17] Z. Zhang, E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas *et al.*, "Mixed linear model approach adapted for genome-wide association studies," *Nature genetics*, vol. 42, no. 4, pp. 355–360, 2010.

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

[19] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[20] J. Zhou and O. G. Troyanskaya, "Predicting effects of non-coding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[21] G. Eraslan, J. Arloth, J. Martins, S. Iurato, D. Czamara, E. B. Binder, F. J. Theis, and N. S. Mueller, "Deepwas: Directly integrating regulatory information into gwas using deep learning supports master regulator mef2c as risk factor for major depressive disorder," *bioRxiv*, p. 069096, 2016.

[22] J. Lanchantin, R. Singh, Z. Lin, and Y. Qi, "Deep motif: Visualizing genomic sequence classifications," *arXiv preprint arXiv:1605.01133*, 2016.

[23] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.

[24] J. Tan, M. Ung, C. Cheng, and C. S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 20. NIH Public Access, 2015, p. 132.

[25] L. Chen, C. Cai, V. Chen, and X. Lu, "Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model," *BMC bioinformatics*, vol. 17, no. 1, p. S9, 2016.

[26] C. A. C. Montañez, P. Fergus, A. Hussain, and D. Al-Jumeily, "Analysis of extremely obese individuals using deep learning stacked autoencoders and genome-wide genetic data," *arXiv preprint arXiv:1804.06262*, 2018.

[27] H. S. Baweja, T. Parhar, O. Mirbod, and S. Nuske, "Stalknet: A deep learning pipeline for high-throughput measurement of plant stalk count and stalk width," in *Field and Service Robotics*. Springer, (accepted) 2017, pp. to–appear.

[28] T. Mueller-Sim, M. Jenkins, J. Abel, and G. Kantor, "The robotanist: A ground-based agricultural robot for high-throughput crop phenotyping," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3634–3639.

[29] E. Hosseini-Asl, R. Keynton, and A. El-Baz, "Alzheimer's disease diagnostics by adaptation of 3d convolutional network," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 126–130.

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