American Journal of Applied Mathematics and Statistics, 2022, Vol. 10, No. 2, 44-51
Available online at http://pubs.sciepub.com/ajams/10/2/2
Published by Science and Education Publishing
DOI:10.12691/ajams-10-2-2

Features Selection in Statistical Classification of High Dimensional Image Derived Maize (*Zea Mays* L.) Phenomic Data

Peter Gachoki[1,*], Moses Muraya[2], Gladys Njoroge[1]

moses.muraya@chuka.ac.ke

[1]Department of Physical Sciences, Chuka University, P.O Box 109-60400, Chuka, Kenya
[2]Department of Plant Sciences, Chuka University, P.O Box 109-60400, Chuka, Kenya
*Corresponding author: pkgachoki@gmail

Received April 17, 2022; Revised May 26, 2022; Accepted June 06, 2022

Abstract Phenotyping has advanced with the application of high throughput phenotyping techniques such automated imaging. This has led to derivation of large quantities of high dimensional phenotypic data that could not have been achieved using manual phenotyping in a single run. Hence, the need for parallel development of statistical techniques that can appropriately handle such large and/or high dimensional data set. Moreover, there is need to come up with a statistical criteria for selecting the best image derived phenotypic features that can be used as best predictors in modelling plant growth. Information on such criteria is limited. The objective of this study is to apply feature importance, feature selection with Shapley values and LASSO regression techniques to find the subset of features with the highest predictive power for subsequent use in modelling maize plant growth using high-dimensional image derived phenotypic data. The study compared the statistical power of these features extraction methods by fitting an XGBoost model using the best features from each selection method. The image derived phenomic data was obtained from Leibniz Institute of Plant Genetics and Crop Plant Research, -Gatersleben, Germany. Data analysis was performed using R-statistical software. The data was subjected to data imputation using k Nearest Neighbours technique. Features extraction was performed using feature importance, Shapley values and LASSO regression. The Shapley values extracted 25 phenotypic features, feature importance extracted 31 features and LASSO regression extracted 12 features. Of the three techniques, the feature importance criterion emerged the best feature selection technique, followed by Shapley values and LASSO regression, respectively. The study demonstrated the potential of using feature importance as a selection technique in reduction of input variables in of high dimensional growth data set.

Keywords: *high throughput phenotyping, high dimensional data, feature extraction, feature importance, Shapley values, LASSO regression*

Cite This Article: Peter Gachoki, Moses Muraya, and Gladys Njoroge, "Features Selection in Statistical Classification of High Dimensional Image Derived Maize (*Zea Mays* L.) Phenomic Data." *American Journal of Applied Mathematics and Statistics*, vol. 10, no. 2 (2022): 44-51. doi: 10.12691/ajams-10-2-2.

1. Introduction

Feature selection is the process of reducing the number of input variables when developing predictive models [1]. Feature selection aims at reducing the computational cost of modelling in addition to improving the performance of the predictive models [1]. Moreover, selection of best phenotypes is a crucial step in breeding for increased crop productivity [2]. In feature selection, the relationship between each input variable and the goal variable is evaluated using statistical techniques, and the input variables with the strongest relationship with the target variable are selected. Although the choice of statistical measures depends on the data type for both the input and output variables, these methods can be quick and successful [1].

Currently there is improvement on phenotyping methods, including use large-scale imaging phenotyping technique, which can derive large and complex data set on all aspect of plant growth and development [3]. This technique allows for collection of many plant features [4]. These features includes image-based projected biomass, dynamic growth, colour, shape descriptors, root and canopy architecture, seed morphology, panicle traits, photosynthetic performance, quantum yield, non-photochemical quenching, leaf disease severity assessments, and leaf health status [4]. This kind of data have complex and many attributes, which may not be appropriately handled by commonly used statistical techniques [5] to full explain plant growth and development. Therefore, the existing techniques may require an enabling data inference technology in addition to algorithm development with the aim of analytically solving complex data sets [6,7]. The main challenge of high throughput derived image data is determination of

discriminating traits that can explain plant growth and development. Therefore, there is need to come up with mathematical algorithms that can optimise the selection of features for such data set. This requires improvement or development of new statistical techniques that can be applied in the selection of a set of predictors.

Many statistical techniques have been used in feature selection including forward selection, backward elimination, recursive feature elimination, univariate selection, and feature importance and correlation matrix with heatmaps [8]. Forward selection, backward elimination and recursive feature elimination works well with small data sets [8]. For high dimensional data set such as the high-throughput image derived data, the selection process would be very slow. This necessitates the use of more robust feature selection methods such as feature importance, feature selection with Shapley values and LASSO regression [9]. However, information on their applicability in such data set is missing. Shapley values technique helps feature selection is performed with ranking-based algorithms [10]. Instead of using the default variable importance, generated by gradient boosting, we select the best features like the ones with the highest Shapley values. The benefit in using Shapley values is clear due to the bias present in native tree-based feature importance [10]. The standard methods tend to overestimate the importance of continuous or high-cardinality categorical variables. This makes it not trustable the importance computation in case of feature shifts or changes in the number of categories [11].

In large data, features selection can also be addressed using the features importance, whereby each feature in the dataset is obtained using the feature importance property of the model, using tree based classifiers [12]. Feature importance gives a score for each feature of the data. In calculation of feature importance, nodes importance is first computed using Gini Importance as;

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)}$$

where ni_{jj} = the importance of node j, w_{jj} = weighted number of samples reaching node j, C_{jj} =the impurity value of node j, $C_{left(j)}$ = child node from left split on node j, $C_{right(jj)}$= child node from right split on node j.

The feature importance is then given as;

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ notes} ni_j}$$

Another feature selection method is LASSO regression. This methods has the ability to nullify the impact of an irrelevant feature in the data [13]. This means that it reduces the coefficient of a feature to zero thus completely eliminating it and hence is better at reducing the variance when the data consists of many insignificant features [14]. This study applied different features selection techniques and compared their discriminating power. Once the most discriminating features are identified, they be used in subsequent development of predictive models for plant growth and development.

2. Methodology

The data was obtained from Leibniz Institute of Plant Genetic and Plant Crop Research (IPK-Gaterleben), Gatersleben, Germany. The data consisted of 252 maize inbred lines cultivated in replicated experiments in a climate-controlled glasshouse. The phenotype data was collected at different developmental time points using high-throughput LemnaTec Phenotyping Platform. The platform uses a set of camera systems to derive 784 image phenotypic features. The analysis of the data involved performing data imputation using k Nearest Neighbours to ensure completeness of the data. Feature extraction was done using feature importance, Shapley values and feature selection using LASSO regression so as to come up with a set of the most discriminating traits. Further, the discriminating power of the various feature extraction techniques was compared by fitting an XGBoost model with the best set of features from each technique.

2.1. *k*-Nearest Neighbours

k-Nearest Neighbours (KNN) implements data imputation by identifying k samples in the dataset that are similar or close in the space. The k samples are then used to estimate the value of the missing data points. Each sample's missing values is imputed using the mean value of the '*k*'-neighbours found in the dataset. The KNN technique uses a Euclidean distance metric to impute the missing values. In the presence of missing coordinates, the Euclidean distance is calculated by ignoring the missing values and scaling up the weight of the non-missing coordinates.

$$d_{xy} = \sqrt{weight \times squared\ \ distance\ from\ present\ \ coordinates}$$

where, $weight = \frac{Total\ number\ of\ \ coordinates}{Number\ of\ \ present\ coordiantes}$.

2.2. Feature Importance Selection Technique

In feature importance, decision trees models, which are ensemble learners, are used to rank the importance of the different features [15]. Feature importance was calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability was calculated by the number of features that reached the node, divided by the total number of features. The higher the value the more important the feature was.

2.3. Shapley Values Feature Selection Technique

This method was used to explain the prediction of a feature, say x, by computing the contribution of each feature to the prediction. Shapley values explained how the prediction was fairly distributed among the features. Shapely values were computes as;

$$\varphi_m(v) = \frac{1}{P} \sum_S \frac{\left[v\left(S \cup \{m\}\right) - v\left(S\right) \right]}{\binom{p-1}{k(S)}}, m = 1,2,3,\ldots,p$$

where $\varphi_m(v)$ was the Shapley value, m = the summation was over all the subsets S of the features $T = \{1, 2, 3, \ldots, p\}$ that were constructed after excluding m. $k(S)$ was the size of S, $v(S)$ was the value achieved by sub set S and $v(S \cup \{m\})$ was the realized value after m joined S.

Essentially, the Shapley value was the average marginal contribution of a feature considering all possible feature combinations.

2.4. Lasso Regression Feature Selection Technique

In LASSO regression, the best features were selected when making predictions on a dataset. This was done by LASSO regression method putting a constraint on the sum of the absolute values of the model parameters. The sum had to be less than a fixed value (upper bound). The method applied a shrinking (regularization) process where it penalized the coefficients of the regression variables, shrinking some of them to zero. The regularization process was controlled by the alpha parameter in the LASSO model. The higher the value of alpha, the higher the chances that the feature coefficients were zero.

The LASSO regression was presented as;

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

Where, λ denoted the amount of shrinkage, λ = 0 implied all features were considered and it was equivalent to the linear regression where only the residual sum of squares was considered to build a predictive model, λ = ∞ implied no feature was considered, that is, as λ closes to infinity eliminated more and more features, the bias increased with increase in λ and variance increased with decrease in λ which was equivalent to; Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients).

2.5. Feature Selection Methods Comparison

The feature selection techniques were compared by fitting an XGBoost models with the features obtained. The various XGBoost models for the various feature selection techniques were compared based on their root mean squared error, mean absolute error and the R-squared metrics.

3. Results and Discussion

3.1. Data Overview and Processing

There were a total of 784 phenotypic features. This demonstrated that image derived phonemics constitute high dimensional data set. Such data may not be of direct use to modelling of plant growth and development due to data complexity. In the case of high dimensional data most classification algorithms cannot be directly applied. This effect on dimensionality is confounded further by intensified noisy and uninformative features. A solution to this bottleneck is to apply features selection techniques to appropriately choose subsets of the explanatory variables. Feature selection have been found to improve the classification accuracy and reduces the risk of over-fitting [1]. These features were selected to come up with a smaller set of phenotypic features that was to be a representative of all the features.

3.2. Data Imputation using *k* Nearest Neighbours

The idea in kNN methods was to identify 'k' samples in the dataset that were similar or close in the space. Then ‘k’ samples were used to estimate the value of the missing data points. Each features’ missing values were imputed using the mean value of the 'k'-neighbours found in the dataset. The mean of the neighbours was taken, or weighted mean, where the distances to neighbours were used as weights. That meant that the closer neighbor was, the more weight it had when taking the mean. The end result in the data imputation ensured completeness in the data points of the features making it ready for the next step of the analysis. A similar study that applied kNN imputation results showed that the KNN imputation method maintained or improved classification accuracy according to most classification algorithms [16]. A survey on missing data in machine learning showed that KNN imputation performed better than the random forest (RF) imputation using RMSE as an evaluation measure on the Iris data on two missingness ratios and the RF performed better than the KNN on the ID fan data on all missingness ratios [17]. This led to a conclusion that, the precision and accuracy of machine learning imputation algorithms depended strongly on the type of data being analyzed, and that there was no clear indication that favored one method over the other [17].

3.3. Features Selection Based on Feature Importance

Based on feature importance: 30 out 784 features were selected at 11 DAS, 30 out 784 features were selected at 13 DAS, 28 out 784 features were selected at 15 DAS, 28 out 784 features were selected at 18 DAS, 28 out 784 features were selected at 20 DAS, 28 out 784 features were selected at 22 DAS, 25 out 784 features were selected at 24 DAS, 26 out 784 features were selected at 26 DAS, 27 out 784 features were selected at 28 DAS, 28 out 784 features were selected at 30 DAS, 27 out 784 features were selected at 32 DAS, 24 out 784 features were selected at 34 DAS, 31 out 784 features were selected at 36 DAS, 29 out 784 features were selected at 38 DAS, 27 out 784 features were selected at 40 DAS and 26 out 784 features were selected at 42 DAS. This feature selection was an indication that the statistical analysis did not have to take all the 784 variables since by feature importance an optimal subset of 31 features (Table 1) could be used as representation of all the features. The feature importance plot for the optimal features selected using feature importance is presented in Figure 1. Feature importance scores played an important role in feature selection which included providing insight into the data and dimensionality reduction. These findings are in

agreement with findings by [18] where feature importance was found as one of the best ways for feature selection in machine learning. A study by [19] also found feature importance as a most popular explanation technique in comparing of feature importance measures as explanations for classification models.

Figure 1. Selected Features as Obtained Using Feature Importance

Table 1. Selected Features as Obtained using Feature Importance

feature	importance .05	importance	importance .95	permutation .error
volume.fluo.prism.norm__mm_3	1.5743632	1.648438	1.670008	2.611369
volume.fluo.area090t.norm__mm_2	1.2633809	1.274378	1.284798	2.018803
top.leaf.length.sum.norm_skeleton	1.1629873	1.262433	1.277481	1.99988
volume.vis.prism.norm__mm_3	1.2127528	1.24389	1.290652	1.970505
top.fluo.area.norm__mm_2	1.1490368	1.164798	1.193612	1.845212
side.fluo.area.norm__mm_2	1.1051697	1.137032	1.141801	1.801227
side.vis.border.length.norm	1.1090012	1.135904	1.146171	1.799439
top.fluo.hull.circumcircle.d.norm	1.0983132	1.102302	1.103506	1.746209
top.vis.area.norm__mm_2	1.0662907	1.099652	1.102573	1.742012
top.vis.hull.pc2.norm	1.0737672	1.089658	1.091134	1.726179
side.fluo.hull.area.norm	1.0720701	1.086806	1.098964	1.721661
side.vis.area.norm.max__mm_2	1.0692625	1.077498	1.094055	1.706916
top.vis.hull.area.norm	1.0496764	1.059866	1.069641	1.678984
side.height.norm__mm_140	1.0463511	1.05916	1.076474	1.677866
volume.fluo.lt.norm__mm_3	1.0500421	1.055616	1.062048	1.672252
top.vis.border.length.norm	1.0419436	1.050145	1.07109	1.663584
volume.vis.lt.norm__mm_3	1.0421839	1.044253	1.05265	1.65425
side.vis.hull.pc2.norm	1.0124049	1.021018	1.026107	1.617444
volume.vis.iap.norm__px_3	1.0100686	1.019535	1.035392	1.615095
side.leaf.length.sum.norm_skeleton	1.0153671	1.019303	1.023861	1.614726
side.leaf.width.average.norm_skeleton	1.0112321	1.018431	1.02451	1.613344
side.fluo.area.norm.max__mm_2	1.011797	1.016131	1.018442	1.609702
side.vis.area.norm__mm_2	1.0104896	1.013084	1.019461	1.604874
top.leaf.width.average.norm_skeleton	1.0113804	1.01185	1.015771	1.60292
side.height.norm__mm_138	1.0012591	1.011101	1.013676	1.601733
volume.fluo.iap.norm__px_3	1.000899	1.008332	1.011192	1.597346
volume.vis.iap_max.norm__px_3	1.0017949	1.004697	1.006748	1.591587
volume.vis.area090t.norm__mm_2	0.9961751	1.003895	1.017454	1.590318
top.vis.hull.pc1.norm	0.998067	1.003548	1.008246	1.589767
side.fluo.border.length.norm	1.0015909	1.003222	1.004157	1.589252
side.fluo.hull.pc2.norm	1.0009482	1.001311	1.002318	1.586225

3.4. Feature Selection using Shapley Values

Table 2. Selected Values as Obtained using Shapley Values

feature	Med	Mean
volume.fluo.prism.norm__mm_3	1.51833	1.51833
side.vis.area.norm.max__mm_2	0.99683	0.99683
side.vis.area.norm__mm_2	0.80765	0.80765
volume.fluo.iap.norm__px_3	0.29418	0.29418
top.leaf.width.average.norm_skeleton	0.28416	0.28416
side.fluo.area.norm.max__mm_2	0.26961	0.26961
side.vis.border.length.norm	0.26854	0.26854
side.fluo.area.norm.min__mm_2	0.26201	0.26201
top.fluo.area.norm__mm_2	0.20775	0.20775
top.fluo.main.axis.normalized.distance.avg	0.18279	0.18279
volume.vis.lt.norm__mm_3	0.16933	0.16933
side.vis.hull.area.norm	0.15019	0.15019
side.fluo.area.norm__mm_2	0.14614	0.14614
side.width.norm__mm	0.1289	0.1289
side.fluo.border.length.norm	0.09324	0.09324
side.height.norm__mm_138	0.07341	0.07341
side.leaf.length.average.norm_skeleton	0.07319	0.07319
side.height.norm__mm_140	0.06551	0.06551
top.vis.border.length.norm	0.03857	0.03857
top.fluo.hull.pc2.norm	0.02583	0.02583
volume.vis.iap_max.norm__px_3	0.02346	0.02346
side.leaf.width.average.norm_skeleton	0.02143	0.02143
side.fluo.hull.pc2.norm	0.01821	0.01821
volume.fluo.iap_max.norm__px_3	0.01373	0.01373
side.vis.hull.pc2.norm	0.01337	0.01337

Based on Shapley values: 25 out 784 features were selected at 11 DAS, 24 out 784 features were selected at 13 DAS, 26 out 784 features were selected at 15 DAS, 24 out 784 features were selected at 18 DAS, 24 out 784 features were selected at 20 DAS, 26 out 784 features were selected at 22 DAS, 22 out 784 features were selected at 24 DAS, 22 out 784 features were selected at 26 DAS, 24 out 784 features were selected at 28 DAS, 25 out 784 features were selected at 30 DAS, 25 out 784 features were selected at 32 DAS, 26 out 784 features were selected at 34 DAS, 25 out 784 features were selected at 36 DAS, 26 out 784 features were selected at 38 DAS, 24 out 784 features were selected at 40 and 24 out 784 features were selected at 42 DAS. The Shapley values technique yielded 25 optimal features as a representation of the 784 features (Table 2). The optimal features were presented diagrammatically in form of a Shap decision plot (Figure 2).

Shapley value of a feature averaged marginal contribution of a feature value across all the possible combinations of features. The computation time increased exponentially with the number of features. The Shap decision plot shows features each contributing to push the model output from the base value (the average model output over the training dataset passed) to the model output. Features pushing the prediction higher were shown in red and those pushing the prediction lower were in blue (Figure 2). The plot sorted the features by the sum of SHAP value magnitudes over all samples and used SHAP

values to show the distribution of the impacts each feature had on the model output. These findings are in agreement with those from a study by [10] that found that feature selection by Shapley values produced a better performing prediction model when compared to feature selection methods such as forward selection and backward elimination. A study by [11] also found feature selection by Shapley values to be a popular feature selection method which was attributed to its solid theoretical foundation.

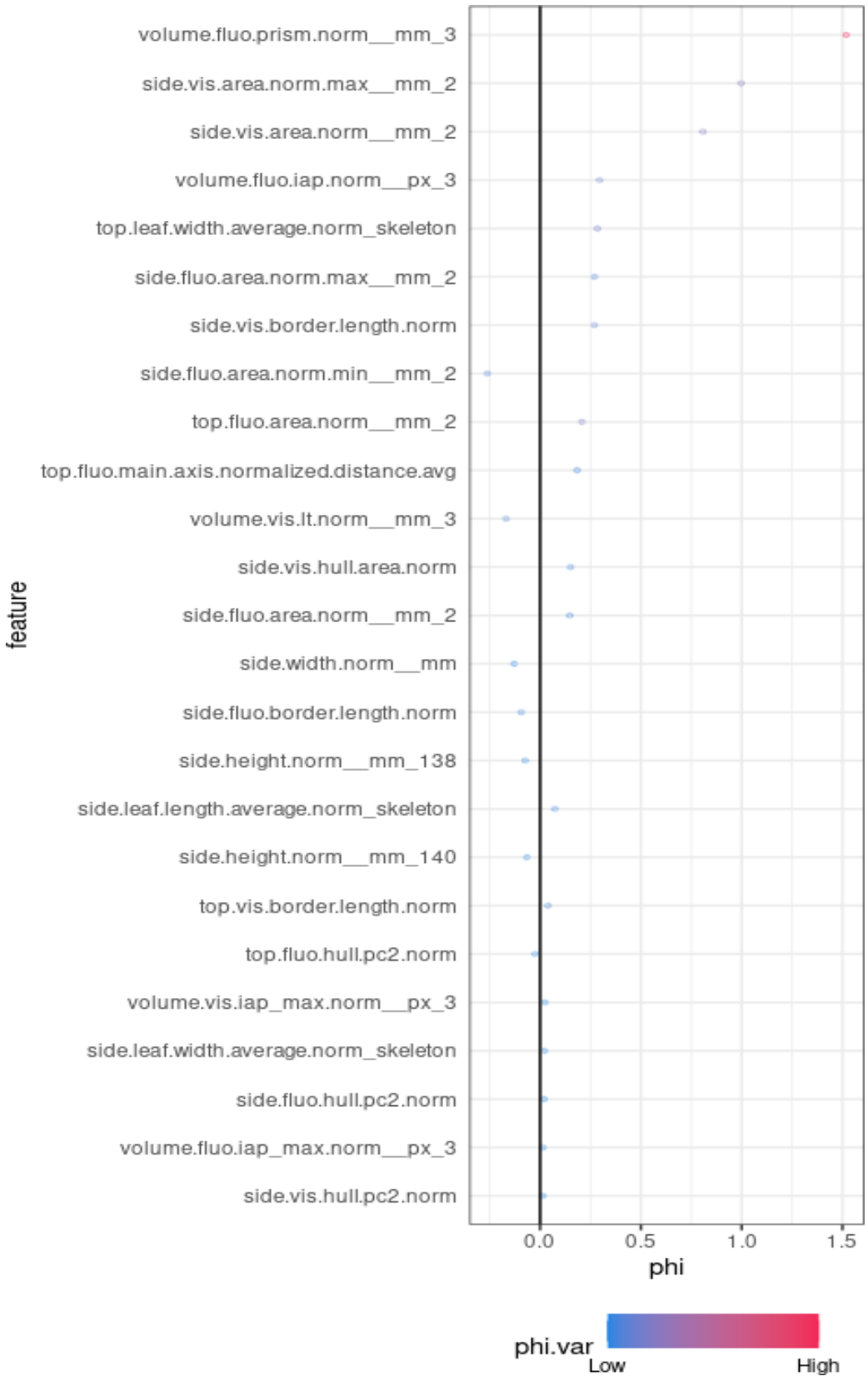

Figure 2. Selected Features as Obtained using Shapley Values

3.5. Feature Selection Using LASSO Regression

Based on LASSO regression: 10 out 784 features were selected at 11 DAS, 30 out 784 features were selected at 13 DAS, 12 out 784 features were selected at 15 DAS, 28 out 784 features were selected at 18 DAS, 31 out 784 features were selected at 20 DAS, 9 out 784 features were selected at 22 DAS, 23 out 784 features were selected at 24 DAS, 26 out 784 features were selected at 26 DAS, 20 out 784 features were selected at 28 DAS, 15 out 784 features were selected at 30 DAS, 14 out 784 features were selected at 32 DAS, 13 out 784 features were selected at 34 DAS, 12 out 784 features were selected at 36 DAS, 18 out 784 features were selected at 38 DAS, 45 out 784 features were selected at 40 DAS and 22 out 784 features were selected at 42 DAS. The optimal features selected using LASSO regression were 12 in number (Table 3). The LASSO regression yielded the least number of optimal features as compared to the feature importance and Shapley values. These findings agree with the findings from a study by [20] who found LASSO regression an important feature selection technique in the domain of high dimensional data. Similar findings were also made by [14] in a study on feature selection technique in predictive modeling for machine learning.

Table 3. Selected Features using LASSO Regression

S/N	Feature
1	side.leaf.length.average.norm_skeleton
2	side.vis.area.norm.max __mm_2
3	side.vis.border.length.norm
4	side.vis.hull.pc2.norm
5	side.width.norm __mm
6	top.fluo.area.norm __mm_2
7	top.fluo.main.axis.normalized.distance.avg
8	top.leaf.width.average.norm_skeleton
9	top.vis.border.length.norm
10	volume.vis.iap_max.norm__px_3
11	side.height.norm__mm_138
12	side.height.norm__mm_140

3.6. Comparison of the Statistical Power of the Feature Selection Techniques

The different feature selection techniques were compared for their statistical power to find the method that yielded the best features. The statistical power of the feature selection techniques was evaluated based on the performance metrics of the extreme gradient model fitted with the obtained set of features. Selection using feature importance was found to provide the best explanatory variables (Table 4). The features obtained using feature importance produced an extreme gradient boosting model that had the lowest value of root mean squared error and thus becoming the best model (Table 4). The model from feature importance features had the highest value of R- squared which meant that the features accounted for the highest amount of variation in the plant biomass (Table 4). Additionally, the model that was based on features selected using feature importance technique had the lowest value of mean absolute error which meant that it was the best model (Table 4). The second best feature selection technique was the Shapley values technique followed by the LASSO regression technique (Table 4). The results showed that feature importance had the best discriminating power in selection of the best features from high dimensional image derived maize (*Zea Mays* L.) phenomic data.

This study results agree with results by [19] which showed that most important features differ depending on the technique used. In a study for feature selection using approximated high-order interaction components of the Shapley value for boosted tree classifier, it was found out that shapely values outperformed other methods in selection of features for evaluating forecasting performance for handling a problem with hundreds of time-lagged input features [21]. Another study that showed similar findings investigated on comparing of feature selection methodology for solving classification problems in finance [22]. The results showed that feature importance technique consistently attributed feature importance, better align with human intuition and better recover influential features. When applied to the Polish bankruptcy data, the method not only discovered all the important features but also produced the correct classifier [22].

Table 4. Statistical Power of the Different Feature Selection Techniques

	All Features			feature importance			Shapley values			LASSO Regression		
day	RMSE	Rsquared	MAE	RMSE	Rsquared	MAE	RMSE	Rsquared	MAE	RMSE	Rsquared	MAE
11	5.07	0.13	3.92	4.69	0.15	3.69	4.80	0.11	3.76	4.78	0.12	3.69
13	4.84	0.16	3.75	4.68	0.15	3.66	4.79	0.11	3.75	4.71	0.14	3.70
15	4.42	0.30	3.45	4.24	0.29	3.23	4.25	0.29	3.23	4.15	0.32	3.15
18	4.49	0.30	3.50	4.21	0.31	3.32	4.29	0.29	3.40	4.26	0.29	3.35
20	4.24	0.36	3.35	4.16	0.32	3.36	4.20	0.31	3.40	4.12	0.33	3.33
22	3.20	0.62	2.48	3.10	0.62	2.44	3.17	0.61	2.49	3.05	0.63	2.40
24	3.11	0.67	2.43	2.86	0.68	2.29	2.87	0.67	2.29	2.90	0.67	2.31
26	2.98	0.68	2.31	2.77	0.70	2.21	2.91	0.66	2.31	2.82	0.68	2.23
28	2.67	0.76	2.08	2.58	0.74	2.08	2.60	0.73	2.10	2.63	0.73	2.14
30	2.63	0.76	2.00	2.43	0.76	1.91	2.50	0.75	1.96	2.48	0.75	1.94
32	2.51	0.78	1.92	2.28	0.79	1.79	2.33	0.78	1.83	2.26	0.79	1.78
34	2.55	0.78	1.96	2.43	0.77	1.92	2.43	0.76	1.93	2.35	0.78	1.84
36	2.31	0.82	1.77	2.16	0.83	1.73	2.19	0.82	1.73	2.20	0.82	1.75
38	2.51	0.80	1.90	2.17	0.81	1.74	2.21	0.80	1.74	2.26	0.80	1.79
40	2.52	0.79	1.91	2.29	0.79	1.76	2.31	0.79	1.79	2.34	0.78	1.82
42	2.59	0.77	1.99	2.37	0.77	1.83	2.36	0.78	1.83	2.36	0.78	1.84

4. Conclusion

In conclusion, the study revealed that in a case where there is a high dimensional data, feature selection by feature importance would be more ideal in coming up with the best features that can be used modelling of plant growth. The feature importance feature selection technique did not only assist in coming up with features that yielded the best growth model but also assisted in immensely reducing the number of features involved in the modelling. The features obtained using feature importance also accounted for the highest variability in plant biomass as compared to the other feature selection methods

Acknowledgments

The authors acknowledge the Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany for making available the data used in this study. The data was obtained with the support of grants from the German Federal Ministry of Education and Research (BMBF) and performed within the ConFed projects (identification numbers: 0315461C). The authors declare no conflicts of interest.

References

[1] Guyon I. and Elisseeff A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research. 1157-1182.

[2] Pieruschka R, Schurr U (2019) Plant phenotyping: past, present, and future. Plant Phenomics: 1-6.

[3] Chen, D., Neumann, K., Friedel, S., Kilian, B., Chen, M., Altmann, T., & Klukas, C. (2014). Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. The Plant Cell, 26(12), 4636-4655.

[4] Klukas, C., Chen, D., & Pape, J. M. (2014). Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. Plant physiology, 165(2), 506-518.

[5] Boyd, D., & Crawford, K. (2011,). Six provocations for big data. In A decade in internet time: Symposium on the dynamics of the internet and society.

[6] Blum, A., Hopcroft, J., & Kannan, R. (2020). Foundations of data science. Cambridge University Press.

[7] Unnisabegum, A., Hussain, M., & Shaik, M. (2019). Data Mining Techniques for Big Data, Vol. 6, Special Issue.

[8] Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1), 3-26.

[9] Duchesnay, E., & Löfstedt, T. (2018). Statistics and Machine Learning in Python. Release 0.1.

[10] Cohen, Shay & Ruppin, Eytan & Dror, Gideon. (2005). Feature Selection Based on the Shapley Value. 665-670.

[11] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. IEEE Access, 9, 144352-144360.

[12] Helwig, N. E. (2017). Data, Covariance, and Correlation Matrix. University of Minnesota (Twin Cities).

[13] Kim, Yongdai & Kim, Jinseog. (2004). Gradient LASSO for feature selection. Proceedings of the 21st International Conference on Machine Learning.

[14] Muthukrishnan, R. & Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. 18-20.

[15] Ghojogh, B., Samad, M. N., Mashhadi, S. A., Kapoor, T., Ali, W., Karray, F., & Crowley, M. (2019). Feature selection and feature extraction in pattern analysis: A literature review. arXiv preprint arXiv:1905.02845.

[16] Pan, R., Yang, T., Cao, J., Lu, K., & Zhang, Z. (2015). Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. Applied Intelligence, 43(3), 614-632.

[17] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8(1), 1-37.

[18] Kaylan, P. (2021). 7 Popular Feature Selection Routines in Machine Learning. https://www.analyticsvidhya.com/blog/2021/03/7-popular-feature-selection-routines-in-machine-learning/.

[19] Saarela, Mirka & Jauhiainen, Susanne. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences. 3.

[20] Giersdorf, J., & Conzelmann, M. (2017). Analysis of feature-selection for LASSO regression models.

[21] Chu, Carlin & Chan, David. (2020). Feature Selection Using Approximated High-Order Interaction Components of the Shapley Value for Boosted Tree Classifier. IEEE Access. PP. 1-1.

[22] Xiaomao, X., Xudong, Z., & Yuanfang, W. (2019, August). A comparison of feature selection methodology for solving classification problems in finance. In Journal of Physics: Conference Series (Vol. 1284, No. 1, p. 012026). IOP Publishing.