JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Machine Learning Approaches to Predict the Abiotic and Biotic Stress Tolerance Genes in Plants-A Survey

Dr.N. Priya[1], A. Amuthavalli[2]

[1]Associate Professor PG Department of Computer ScienceS.D.N.B. Vaishnav College for WomenUniversity of MadrasChennai, Tamil Nadu, India

[2]Assistant Professor Department of Computer ScienceHindustan College of Arts & Science University of MadrasChennai, Tamil Nadu, India

drnpriya2015@gmail.com[1], amuthavalli@yahoo.com[2]

Received: 05 May 2020 Revised: and Accepted: 15 July 2020

ABSTRACT: Abiotic and biotic stresses, as a part of the normal ecology, seriously impact crop productivity. Environmental pollution, climate change and global warming increase the impact of biotic and abiotic stresses on plant growth and productivity worldwide. The intensification in life-threatening climate change, continuous exposure of pesticides and pathogens warrants developing crop varieties that can tolerate multiple stresses. Thus, finding genes involved in resisting stress conditions is an agronomic importance to improve crop production and supply. With the vast genomic sequence data available from public data base and huge amount of data on expressed genes from various plants, computational approaches to find genes associated with the stress tolerance has become indispensable. Traditional databases are not designed to identification and retrieval of genes especially triggered when plants exposed to various stress factors. Recently, several machine learning approaches have been developed to identify and predict genes with agronomic traits. Different Clustering and classification methods and tools were used to predict the abiotic and biotic stress tolerance genes for plants. The objective of the paper is to discuss currently developed machine learning methods applied in plant biological data, particularly for predicting abiotic and biotic stress related genes to improve crop production and supplying food to an ever-increasing population.

KEYWORDS: Machine Learning, Abiotic Stress, Biotic Stress, Prediction, SVM, Bayesian, Random forest.

I. INTRODUCTION

Plants as sessile organisms are exposed continuously a several types of stress conditions. Stress factors can be divided into abiotic stress triggered by drought, salinity, UV radiation, cold, wounding and biotic stress caused by nematodes, fungi, bacteria and insects [1]. The existence of both abiotic and biotic stresses has a vast impact in crop yield and economic losses experienced by agriculture worldwide. Development of new varieties of crop plants tolerance to specific abiotic or biotic stress is the need of the moment to compact ever growing world population and rising food needs.

Plants have evolved with unique defence mechanisms at molecular and cellular level to combat various biotic and abiotic stress conditions. Under stress conditions, plant's immune system is get activated through various molecular level processes such as increase or decrease in gene expression related to transcription factors, signal transduction and kinase cascade pathways, hormone signalling and heat shock proteins provides immunity to plants [2]. Identification of differential expression pattern of genes under biotic and abiotic condition leads to improved knowledge on plant stress mechanisms [3]. Recent advancements in high throughput technologies and availability of large plant sequence data provide an opportunity to identify genes within different stress conditions.

Bioinformatics tools are effectively used in plants for the analysis and classification of genes [4]. Several bioinformatics studies have analysed microarray samples of gene expression under different conditions for the same species however these methods are not suitable for identifying homologous genes from tropical plants

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

[5].Machine learning (ML) is a field of computer science that uses algorithms and existing samples to capture characteristics of target patterns is a novel strategy for the prediction and classification of gene function based on integration of multiple data sources and ideal for interpreting unannotated deluge amount of data. Therefore development of effective algorithms is warranted for organizing and retrieving gene sequences.

II. PLANTS BIOTIC AND/OR ABIOTIC STRESS

Observing the principles of abiotic and biotic stress responses, tolerance and adaptation remains important in plant physiology research to develop better varieties of crop plants. The analyse of the functions of stress-inducible genes is an important to know the molecular mechanisms of stress tolerance, the responses of plants and also to improve the stress tolerance of plants yields by gene manipulation. The stress tolerance genes involved in biotic and abiotic stresses listed in Table 1

III. LIST OF STRESS TOLERANCE GENES

Table 1. Genes involved in Abiotic & Biotic Stres

Reference Genes for stress Response	Stress Conditions
Alpha tubulin	Sulphate starvation, salt, drought, Bacterial infections
Actin 2	cold, salinity, oxidative, heat stock, high in radiation
Actin 3	Salt stress, drought stress,
Actin 2/8	Salt, mannitol, drought and cold
Ubiquitin-conjugating enzyme 28	Salinity, high and low temperature stress, UV radiation
Ubiquitin-conjugating enzyme 5	Drought, mannitol and salt, Fungal infections
Tubulin2	Sucrose, NaCl, mannitol, paclobutrazol hormonal
Elongation factoir 1-α	Metal stress
Pentatricopeptide-repeat proteins	Metal stress
Elongation factor-1a	Metal stress
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	Salt, mannitol, drought, cold and bacterial infections
Polyubiquitin	Multiple stress
Initiation factor	Cold, Drought
Transcription factors	Multiple stress
Heat Shock proteins	Multiple stress
early responsive to dehydration protein (ERD15)	Drought
Clathrin adaptor complex proteins	Cold, Drought
Cyclophilin	Drought
F-box/kelch-repeat protein	Plant viruses
Ribosomal protein L2	salt, drought
Galactinol synthase	Cold

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Choline kinase	High temperature
Osmotin	High temperature
Dehydrin	Cold
glycine rich protein	Cold
Aquaporin	Freezing
Remorin	Drought
Cysteine synthase	Salt
Calretic	Ozone
Uridylate kinase	Plant viruses
Ubiquitin-conjugating enzyme 3	Plant viruses

IV. COMPUTATIONAL PREDICTION TOOLS FOR ABIOTIC/BIOTIC STRESS GENES

Identification of genes in genomic DNA sequences by computational methods is a problem that has attracted considerable research. The machine learning algorithms like supervised and unsupervised have been used for the identification of genes and non-coding DNA and database homology searching to predict the gene structures in genomic sequences. Supervised learning (SL) is based on labeled data output based on training data set whereas unsupervised learning is based on unlabeled data [6]. The choice of applying specific tool is depends on different tasks and data sets. Based on discrete or continuous output, SL algorithms can be further classified into support vector machine (SVM), random forest and bayesian network. Unsupervised learning algorithms can be classified like Hierarchical clustering (HCL), K-means clustering (KMC) and Principal Component Analysis (PCA) which are primarily used for analysing unbalanced data (sample with less significant amount of data). The Types of machine learning methods are depicted in Figure 1.

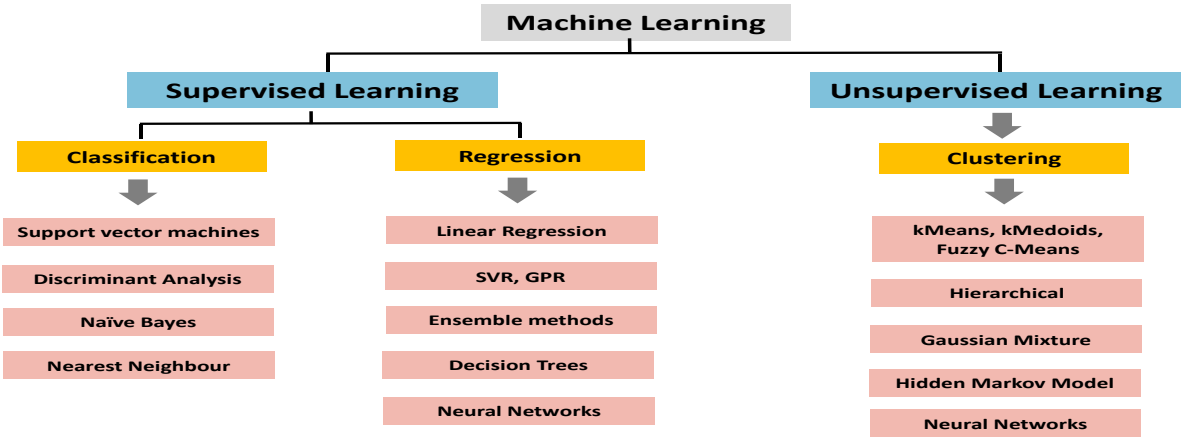


Figure 1. Machine Learning Methods

2.1. Support Vector Machines

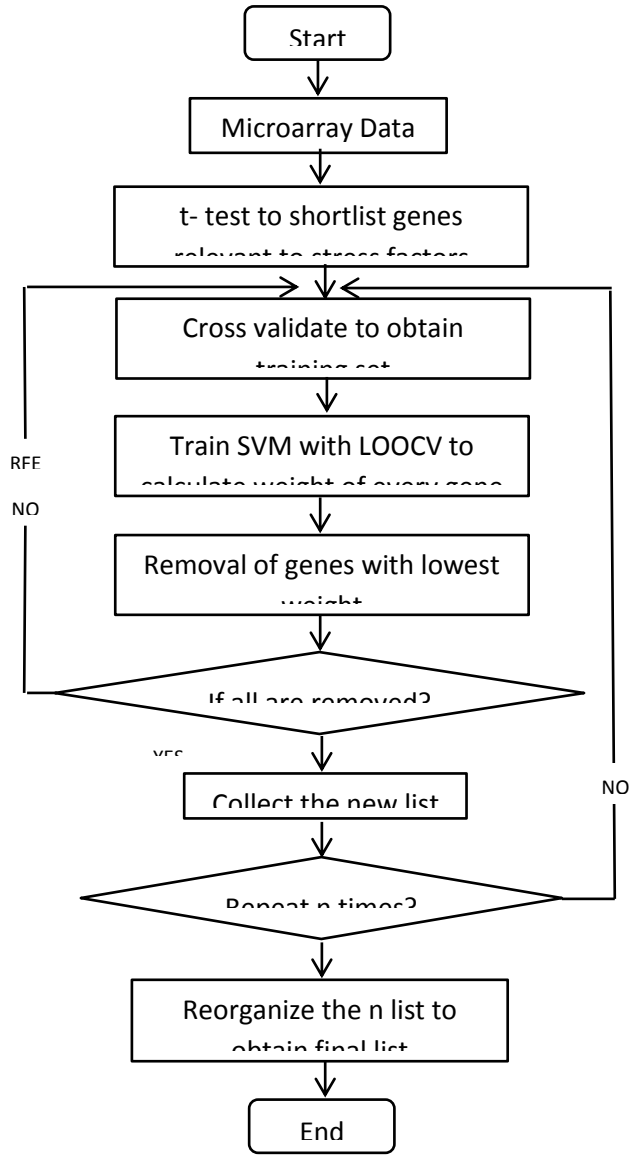
Molecular level understanding of biological conditions is commonly evaluated by identification of genes that are differentially expressed (DE) under normal versus stress situations. Methodologies such as microarray or RNA-sequence profiling are routinely used to get sequence data on differentially expressed genes (DEGs). However, data generated with these methodologies are loaded with high dimensionality of feature space i.e. availability of too many gene or sequence variables for very limited samples. Machine learning algorithms such as SVM is widely applied to classify more classes of data types and predict plant genes expressed in abiotic and biotic stress conditions. SVM classify two or more classes of binary data types to identify the correct label for unknown datasets. SVM precisely segregate positively and negatively labeled datasets by constructing the models using an optimal hyper-plane either through linear or nonlinear based on kernel types. The iterative

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

training algorithm is used to minimize an error function in SVM and adopts optimization parameters and kernel functions such as sigmoid, linear, polynomial, and radial basis function [7]. Additional feature selection for obtaining more accurately classified top features are selected through recursive-support vector machines (R-SVM).

Rafi Shaik et al.[8] reported that the R-SVM for analysing and classifying 559 microarray data of *Oryza sativa* (rice) indicated in 7 abiotic stress conditions like drought, cold, thermal shock, metal toxicity, osmotic, salt, nutrient and 6 biotic stresses like fungal and bacterial infection, insect, nematode, virus and weed infestation. In this study, specific stress responsive genes that control interaction between abiotic and biotic stresses are predicted from 1377 common DEGs. R-SVM identified 540 genes involved in stress condition with 100% accuracy and with LOOCV (leave one out cross validation) method, 88 genes are predicted with 95% accuracy.

Identification of genes involved in environmental stress especially resisting to drought conditions has high agronomic importance. The Support Vector Machine-Recursive Feature Elimination (SVM-RFE) as a feature selection method, Liang et al.[9] developed a tool for predicting genes known to be represented in drought resistance from *Arabidopsis thaliana.* In this method, a t-test with p-value threshold as 0.001 was employed to filter out genes non-related to biotic and abiotic stress factors and finally 736 genes are shortlisted related to drought resistance. In order to get high prediction accuracy many rounds of bootstrapping and sequence backward selection (SBS) procedure in the SVM-RFE was carried out. To handle limited data and to generate training data sets a LOOCV methodology is adopted for training SVM. Out of 736 genes screened, 10 genes are involved in water tolerance, in which 7 of them are belongs to drought resistance. The common algorithm methodology for feature selection used in the SVM-RFE is represented in Figure 2.



ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Wang et al. [10] identified salt tolerance genes from *Oryza sativa* (Rice) microarray data using SVM-RFE. The linear kernel function is used to construct the linear SVM and t-test as a preliminary screening on features from 57381 genes to shortlist 541 genes with p-value less than 1%. Additional features like LOOCV is applied to handle limited data sets and cross validated to get a top ranked 10 genes related salt tolerance.

Sandeep et al. [11] developed a SVM tool NBSPred to differentiate nucleotide binding site leucine-rich repeats (NBSLRRs) involved in plant defence proteins (R-protein). SVM light package is used as SVM classifiers to predict NBSLRR proteins. Five-fold cross validation technique is applied to identify best classifiers with both positive and negative datasets which are further divided into five parts each having almost equal number of proteins. A Pfam domain based selection of R-protein sequences is performed for both positive and negative datasets [12]. A protein sequences totally 974 are considered as training set, which are selected based on 6 composition-based amino-acid characteristics (amino acid, dipeptide, tripeptide, multiplet, charge and hydrophobicity composition). The threshold dependent parameters sensitivity (SN), specificity (SP) and Mathew's correlation coefficient (MCC) are used to calculate prediction performance of NBSPred. Out of 588 models built from training datasets, 132 models are finally selected with mean MCC greater than 0.95 score and with 91.12% accuracy.

2.2 Random Forest Approach

Random Forest (RF) is a supervised machine learning method based on decision tree based algorithm which selects random subsets of features and makes class prediction based on information gain. RF based classification is fast in prediction especially problems involved in analysing big data sets. RF is commonly used in classification and identification of abiotic and biotic stress genes. Proteins which commonly involved in three-dimensional (3D) domain swapping are related with different abiotic and biotic stress conditions. RF to predict protein sequences that belongs to 3D-domain swapping in *Ocimum tenuiflorum*. In this method, Physiochemical features of 439 sequences involved in 3D-domain swapping are created from amino acid index data base (AAINDEX) and WEKA is implanted to shortlist best features. Around 25% of Ocimum genome is predicted to be in 3D-domain swapping of which 12% (1158 sequences) of predicted protein sequences are regulated in abiotic stress [13]. Random Forest approach accurately classified genes involved in biotic and abiotic stresses conditions. Classification of different stress conditions from 1377 DEGs are evaluated through normalized and pareto scaled intensities. Measurement of prediction error in Random Forest is natively done through out-of-bag estimate (OOB) and it is claimed be unbiased in the dataset. RF classified 8 of the 10 biotic and abiotic stresses from rice DEGs with 100% accuracy.

Antifreeze proteins (AFPs) provide tolerance for overwintering plants against antifreeze stress. Therefore identification of AFPs may give key information on mechanism of action of AFP's on interaction with ice-binding crystals and will enable application of AFPs in agriculture for boosting yield of crops in cooler climates [14]. Various RF based ensemble classification has been used for deciphering AFPs. Quality of prediction for AFPs and handling the unbalancing in data is assessed using a statistical parameter called MCC ranges from -1 to 1 in which MCC =1 is the greatest possible prediction value whereas MCC = -1 is the lowest possible prediction. Other statistical parameters like accuracy, sensitivity, specificity and Youden's index also used for validating the prediction performance of the machine learning tools.

Yang et al. [15] proposed "AFP-Ensemble" web server based tool for prediction of AFPs from protein sequences with maximum sensitivity (MCC = 0.892), accuracy (MCC = 0.938) and specificity (MCC = 0.940). A study used a plant seed protein sequences as a data set and developed classification models created on discriminatory capabilities of hybrid feature spaces and a feature vectors are mapped with protein sequences and features like physicochemical properties, domain characteristics, sequence composition and evolutionary information of proteins are considered. The proportion between the numbers of negative samples with positive samples after the random sampling is defined as G parameter and its value has significant impact on AFPs prediction. The parameter G=9 is selected to get higher sensitivity, accuracy and balanced accuracy using the negative and positive sample training sets. The analysis of variance and incremental feature selection (ANOVA-IFS) method is employed using the optimal parameter G and to select high discerning features for shortlisting AFPs.

Kandaswamy et al. [16] reported a ''AFP-Pred'' tool developed based on RF approach for predicting AFPs. For positive and negative training sets protein sequences are randomly selected based on 300 plant seed antifreeze and non-antifreeze domains from 9493 non antifreeze proteins and 481 antifreeze proteins respectively. Protein sequences are determined by 10 prominent features selected based on ReliefF feature subset selection method. In order to exclude the memory effects during the cross validation process a statistical parameter jackknife method is used in ''AFP-Pred' for effectively predicting AFPs.

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Table 2 . Performance comparisons of machine learning tools developed using random forest classifier

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
AFP-Ensemble	89.2	94.0	93.8
AFP-Pred	91.16	77.04	77.34
Cryoprotect	87.27	88.30	88.28

AFP-Pred obtained sensitivity (MCC = 0.847), specificity (MCC = 0.840), accuracy (MCC = 0.843) and balanced accuracy (MCC = 0.844). Especially prediction accuracy of random forest based ‘‘AFP-Pred’’ was 6 - 7% higher than other classifiers like K-nearest neighbor and Naıve Bayes methods.

Reny et al.[17] developed a random forest based webserver tool “Cryoprotect’’ capable of predicting AFPs from large data set sequence of non AFPs and AFPs are 9139 and 478 respectively . “Cryoprotect” methodology adopted interpretable features like dipeptide composition, amino acid composition and physicochemical properties for characterization of AFPs using propensity score analysis with prediction accuracy of 88.28%. The performance of comparisons using RF classifier is shown in Table 2.

2.3 Bayesian Networks

Bayesian network (BN) is a probabilistic graphical model widely used to analyse the biological associations and the probabilistic environment of biological pathways. BN addresses the problem of detecting gene differences based on any one of the linear regression or Hidden Markov Models or multivariate auto-regressive models. Recently, Dynamic Bayesian Network (DBN) models have been developed with correlation, partial correlation to handle large samples and efficiently predict time dependent gene expression levels [18].

Lahiri et al. [19] applied BN with a utility based inference algorithm and modeled transcription factor activation pathway genes specifically WRKY18 known to be regulated in drought stress from the model plant, *Arabidopsis thaliana*. The gene and transcription factors of microarray expression data in plant (Abscisic acid) activated WRKY transcription factor network and applied the utility-based inference technique to find the important regulators of drought stress response genes. The analysis exposed that the WRKY18 is the highest of activated drought response and WRKY60 had the second inducing drought response. Activation or inhibition of the node is associated with conditional probability, which in turn determined by the parameters characteristic of the real biological processes or local probabilities inferred from the given data shown in Figure 3. Netica and R-software are applied to carry out utility calculations and stimulations respectively. In order to handle limited number of data points, concept of utilities is applied in BN for deducing the best node for intervention and measuring the efficacy of node.

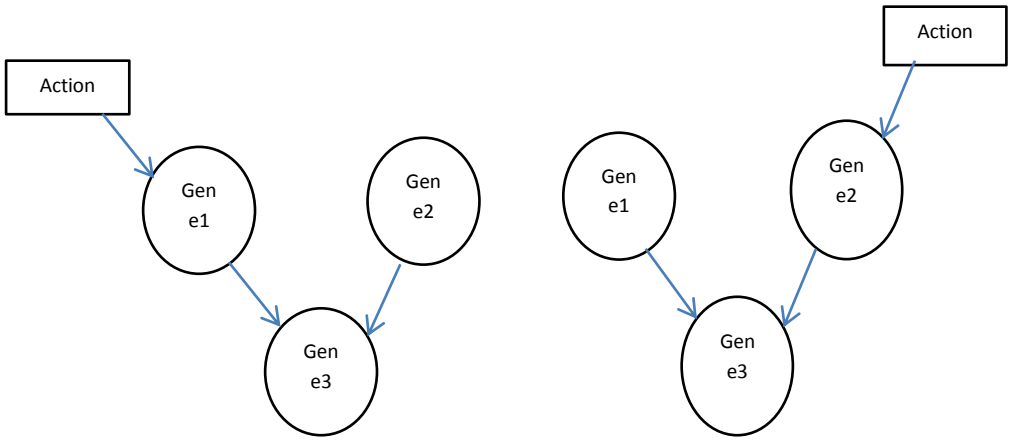


Figure 3. BN describes the probabilistic relationships that exist among gene1, gene2 and gene3

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

Lin at al. [20] developed a Bayesian-based method for analysing transcriptional data and established an association between phenotypic variation of a *Zea mays* plant (Maize) and expression patterns of transcriptional factors (TF) associated with biotic and abiotic stress. Bayesian-based genome-wide association studies (GWAS) predicted genes related to 13 traits of 369 diverse Maize inbreds lines. The software GenSel v4.1 [21] is used to construct Bayesian-based GWAS and its accuracy is estimated through tenfold cross validation conducted using R "cvTools" [22].

Lauro et al.[23] using Bayesian analysis identified family of Hybrid proline-rich proteins (HyPRPs) from Glycine max (Soya bean) induced in response to biotic stress caused by Asian soybean rust disease. Software tool MrBayes v.3.1.2 [24] developed with Bayesian based algorithm is applied to carry out cluster analysis using microarray and Real time PCR data from different soyabean plant tissues to predict putative HyPRPs proteins and their expression patterns.

Wong et al.[25] using the combination of empirical Bayesian and frequentist theory to identify stress-associated genes from the model plant *Thellungiella salsuginea.* A cDNA microarray data contains 3,628 unique genes are obtained from *Thellungiella* plant libraries which are subjected to various stresses, Microarray data is normalized with the Joint Lowess method to avoid any spatial bias and intensity-dependent concurrently. After normalization, differentially regulated genes under control and treatments (different stress conditions) are predicted using a Bayesian model. A total of 149 genes are predicted to be regulated in cold , salinity and drought.

Michael et al. [26] used agglomerative hierarchical clustering algorithm in BN model and identified Cysteine-rich Receptor-like Kinase (CRK) genes from *Arabidopsis thaliana* which are activated under stress response. The p-value and the false discovery rate correction of normalized sequence data are carried out through SPSS tool and Benjamini Hochberg procedures respectively. Bayesian hierarchical clustering combined with parametric bootstrapping is applied for handling sample variation and to numerically measure similarity between control genes versus various treatment conditions. A total of 44 CRKs are screened for transcriptional expression patterns in *Arabidopsis thaliana* plants exposed with ozone (O3), high intensity light and pathogen. A total of 25 CRKs are up regulated in response to O3 and pathogen attack whereas no CRKs are induced upon light dependent treatment. In plants, hierarchical clustering has been widely applied to find the genes involved in abiotic and biotic stresses [27]. Hierarchical clustering based analysis gives a colour representation for each of primary data with a graphical output and in turn allows an intuitive understanding of data either quantitatively or qualitatively [28].

Campos et al. [29] predicted pathogenesis-related proteins (PR proteins) accumulated in *Citrus sinensis* (Citrus) plant in response to pathological or related situations are evaluated using HCL method. In citrus EST database, a total of 3,103 putative transcripts representing PR protein families were deciphered**.** Overall 17 *PR*-like genes are highly expressed in either with pathogens or drought-stressed conditions.

Taji et al. [30] performed comparative cDNA microarray analyses of model plants *Arabidopsis thaliana* and *Thellungiella halophile* (Salt cress) activated under several oxidative stress conditions. HCL is implemented to analyze differential gene expression pattern which indicated various genes are highly expressed in *Arabidopsis* fashioned by abiotic and biotic stresses but the same genes are normally expressed in salt cress in unstressed conditions. Similarly HCL is applied for analysing differentially expressed abiotic and biotic related genes like calcineurin B-like (CBL) protein kinases from pineapple [31] Germin-Like Protein Family from rice and Arabidopsis [32], aquaporins from Canola [33].

2.4 K-means Clustering (KMC)

K-means is one of unsupervised iterative algorithm works well for solving clustering problem. This involves classifying a set of data into a certain number of clusters based on similarity index in which k is an integer [34]. Moore at al. [35] utilized KMC as one of unsupervised algorithm to compare the pattern of gene expression in stress and normal plant growing conditions

Zhang et al. [36] identified critical genes involved in drought tolerance in the plant *Salix psammophila* (desert willow). KMC based clustering is carried out for 8172 differentially expressed genes obtained from root transcriptomic data. Expression pattern of reference genes involved in in several stress related conditions are classified into nine clusters and compared. A total of 672 transcription factors derived form 45 gene families are predicted to be specifically expressed in drought conditions.

Uygun et al. [37] reported that k-means based clustering analysis is highly informative for analysing functional association of biotic and biotic genes. Around 22,000 genes from Arabidopsis microarray datasets related to various stress factors is used for clustering analysis. Out of 225 pathways analysed, 95% of genes comprise the

same pathway not co expressed and merely the 5% of genes are identified to have higher co expression ratio under given stress conditions.

2.5 Principal Component Analysis

PCA is one of the statistical methods broadly applied to classify genetic variation and phenotypic plant traits based on similarities. Olivas et al.[38] evaluated transcriptomic changes in Arabidopsis plant stressed either by herbivore attack or combination of herbivore and fungal pathogen attack using PCA analysis. The regularized log2-transformed data is created with software SIMCA P+ v.12 (Umetrics, Umea, Sweden) and the PCA in the DESEQ2 package in R [39] is applied to identify 915 genes that are DE.

Verma et al. [40] using PCA methodology studied genetic variation, diversity of rice varieties related to abiotic and biotic stress tolerance. PCA using R packages (http://www.R-project.org/) is applied to analyse the root and drought tolerant traits of 114 rice genotype data. PCA revealed 147 alleles in which genetic diversity is mainly contributed by differences in various stages of plant growth.

Razzaq et al. [41] applied PCA to screen drought tolerant genotypes of *Helianthus annuus* (Sunflower) plant. Stress index parameters such as fresh weight stress tolerance index, promptness index, seedling height stress tolerance index, germination stress tolerance index and dry matter stress tolerance index (DMSI) is computed from data obtained from plants subjected to drought and normal treatments. After the analysis of variance PCA analysis was applied for the stress index data. Among 60 genotypes tested, 10 are appeared as drought tolerant and 6 are observed as drought sensitive.

Tahmasebi et al.[42] applied PCA methodology to evaluate large scale meta-analysis of microarray data obtained from *Gossypium* spp. (Cotton plant) subjected to diverse abiotic stress conditions. SVA R tool based on empirical Bayes method [43] implemented to correct any batch effects through meta-analysis outcome. Differentially expressed gene with p=value <0.01 are considered as potential ones either down regulated or up regulated under abiotic stress.

2.6 Other Techniques

The major tasks in machine learning are dealing high-dimension (large number of genes) and low-sample data issues. Recently Kang et al. [44] developed a model 'StressGenePred' is designed based on Confident Multiple Choice Learning (CMCL) loss, a twin neural network model and feature embedding method. The tool is constructed based on twin classification models in which includes logical layer is used as symmetrical way with belongs to input and output. The CMCL loss is implemented to create twin model on selecting genes precise to single stress conditions. StressGenePred precisely classified differentially expressed genes from Arabidopsis subjected to drought, salt, cold and heat stresses.

Yue et al. [45] constructed crops stress tolerance database (CSTB) based on convolutional neural network technology to catalogue stress related genes and proteins from various crop species. The collection of 33,207 sequence data from rice, barely, wheat and maize plants are used to develop a processed word vector which in turn embedded into the Embedding layer of TextCNN [46] and used as test set to train the trained model. Classification of data set uncovered total of 1821 genes out of 1,371 is related to biotic stress and 207 are related to abiotic stress.

V. CONCLUSION

In summary, machine learning algorithms are able to predict properties for a new data. The algorithms decision trees, PCA, SVM, RF and BN used to provide a technique to efficiently classify the gene data and predict stress tolerance genes for plants to improve the crop productivity for ever growing population. The evaluation of machine learning algorithms is subject to the versatility of the dataset used, credibility of the different feature extraction methods, the performance of different classifiers and the combination of the classifiers etc. It is a well-known fact that no much work has been reported in the computational prediction of stress linked genes in tropical plants and hence more research is needed in this area to help the biologists to annotate the newly sequenced genes.

VI. REFERENCES

[1] Piasecka A, Kachlicki P, Stobiecki M. Analytical Methods for Detection of Plant Metabolomes Changes in Response to Biotic and Abiotic Stresses. Int J Mol Sci. 2019 Jan 17;20(2):379. doi: 10.3390/ijms20020379. PubMed PMID: 30658398; PubMed Central PMCID: PMC6358739.

[2] Sinha AK, Jaggi M, Raghuram B, Tuteja N. Mitogen-activated protein kinase signaling in plants under abiotic stress. Plant Signal Behav. 2011;6(2):196–203. oi:10.4161/psb.6.2.14701

[3] Bidhan Roy, S.K. Noren, Asit B. Mandal and Asit K. Basu, 2011. Genetic Engineering for Abiotic Stress Tolerance in Agricultural Crops. Biotechnology, 10: 1-22.

[4] Ong Q, Nguyen P, Thao NP, Le L. Bioinformatics Approach in Plant Genomic Research. Curr Genomics. 2016 Aug;17(4):368-78. doi: 0.2174/1389202917666160331202956. PubMed PMID: 27499685; PubMed Central PMCID: PMC4955030.

[5] Lin H. Microarray data analysis of gene expression evolution. Gene Regul Syst Bio. 2009 Nov 27;3:211-4. PubMed PMID: 20054404; PubMed Central PMCID: PMC2796969.

[6] Shanwen Sun, Chunyu Wang, Hui Ding, Quan Zou.Machine learning and its applications in plant molecular studies, *Briefings in Functional Genomics*, Volume 19, Issue 1, January 2020, Pages 40–48

[7] Damasevicius, Robertas. (2010). Optimization of SVM parameters for recognition of regulatory DNA sequences. TOP. 18. 339-353. 10.1007/s11750-010-0152-x.

[8] Rafi Shaik, Wusirika Ramakrishna Machine learning approaches distinguish multiple stress conditions using stress responsive genes and identify candidate genes for broad resistance in rice Published November 2013. DOI: https://doi.org/10.1104/pp.113.22586

[9] Yanchan Liang,Fan Zhang,Juexin Wang,Trupti Joshi,Yan Wang,Dong Xu Prediction Of Drought Resistance Genes in Arabidopsis Thaliana Using SVM-RFE , PLoS ONE 6(7):21750.doi:10.1371/jpornal.pone.0021750.

[10] Juexin Wang, Fan Zhang Identification of Salt Tolerance Genes in Rice from Microarray Data using SVM-RFE,Published in BICoB 2011 Biology, Computer Science

[11] Sandeep K. Kushwaha, Pallavi Chauhan, Katarina Hedlund, Dag Ahrén, NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR

[12] prediction, Bioinformatics, Volume 32, Issue 8, 15 April 2016, Pages 1223–2225, https://doi.org/10.1093/bioinformatics/btv714

[13] Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Res. 2008 Jan;36(Database issue):D281-8. Epub 2007 Nov 26.

[14] Upadhyay AK, Sowdhamini R. Genome-Wide Analysis of Domain-Swap Predicted Products in the Genome of Anti-Stress Medicinal Plant: *Ocimum tenuiflorum*. *Bioinform Biol Insights*. 2019;13:1177932218821362. Published 2019 Jan 9. doi:10.1177/1177932218821362

[15] Gupta R, Deswal R.Antifreeze proteins enable plants to survive in freezing conditions. J Biosci. 2014 Dec;39(5):931-44.

[16] Yang R, Zhang C, Gao R, Zhang L. An Effective Antifreeze Protein Predictor with Ensemble Classifiers and Comprehensive Sequence Descriptors. *Int J Mol Sci*. 2015;16(9):21191–21214. Published 2015 Sep 7. doi:10.3390/ijms160921191

[17] Kandaswamy KK. Chou KC, Martinetz T, Möller S, Suganthan PN, Sridharan S, Pugalenthi G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol. 2011 Feb 7;270(1):56-62. doi: 10.1016/j.jtbi.2010.10.037. Epub 2010 Nov 4.

[18] Reny Pratiwi, Aijaz Ahmad Malik, Nalini Schaduangrat, Virapong Prachayasittikul, Jarl E.S. Wikberg, Chanin Nantasenamat, Watshara Shoombuatong CryoProtect: A Web Server for Classifying Antifreeze Proteins from Nonantifreeze Proteins Volume 2017 Article ID 9861752 https://doi.org/10.1155/2017/9861752

[19] Savage RS, Heller K, Xu Y, Ghahramani Z, Truman WM, Grant M, Denby KJ, Wild DL. R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* 10, 242 (2009). https://doi.org/10.1186/1471-2105-10-242

[20] Lahiri, A., Venkatasubramani, P. & Datta, A. Bayesian modeling of plant drought resistance pathway. *BMC Plant Biol* 19, 96 (2019). https://doi.org/10.1186/s12870-019-1684-3

[21] Lin, H., Liu, Q., Li, X. *et al.* Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. *Genome Biol* 18, 192 (2017). https://doi.org/10.1186/s13059-017-1328-6

[22] Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186.

[23] Alfons A. cvTools: cross-validation tools for regression models. R package version 03. 2012;2

[24] Lauro Bücker Neto , Rafael Rodrigues de Oliveira , Beatriz Wiebke-Strohm , Marta Bencke , Ricardo Luís Mayer Weber , Caroline Cabreira , Ricardo Vilela Abdelnoor , Francismar Correa Marcelino , Maria Helena Bodanese Zanettini and Luciane Maria Pereira Passaglia, Identification of the soybean HyPRP family and specific gene response to Asian soybean rust disease ,Genetics and Molecular Biology, 36, 2, 214-224 (2013)

[25] Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755

[26] Chui E. Wong, Yong Li, Aurelie Labbe, David Guevara, Paulo Nuin, Brett Whitty, Claudia Diaz, G. Brian Golding, Gordon R. Gray, Elizabeth A. Weretilnyk, Marilyn Griffith, Barbara A. Moffatt Transcriptional Profiling Implicates Novel Interactions between Abiotic Stress and Hormonal Responses in Thellungiella, a Close Relative of Arabidopsis Plant Physiology Vol. 140, No. 4 (Apr., 2006), pp. 1437-1450

[27] Michael Wrzaczek, Mikael Brosché, Jarkko Salojärvi, Saijaliisa Kangasjärvi, Niina Idänheimo, Sophia Mersmann, Silke Robatzek, Stanisław Karpiński, Barbara Karpińska & Jaakko Kangasjärvi Transcriptional regulation of the CRK/DUF26 group of Receptor-like protein kinases by ozone and plant hormones in Arabidopsis. *BMC Plant Biol* 10, 95 (2010). https://doi.org/10.1186/1471-2229-10-95

[28] Simon Rasmussen, Pankaj Barah, Maria Cristina Suarez-Rodriguez, Simon Bressendorff, Pia Friis, Paolo Costantino, Atle M. Bones, Henrik Bjørn Nielsen, John Mundy Transcriptome Responses to Combinations of Stresses in Arabidopsis .Published April 2013. DOI: https://doi.org/10.1104/pp.112.210773

[29] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998 Dec;95(25):14863-14868. DOI: 10.1073/pnas.95.25.14863.

[30] Campos, Magnólia A., Rosa, Daniel D., Teixeira, Juliana Érika C., Targon, Maria Luisa P.N., Souza, Alessandra A., Paiva, Luciano V., Stach-Machado, Dagmar R., Machado, Marcos A PR gene families of citrus: their organ specific-biotic and abiotic inducible expression profiles based on ESTs approach Genetics and Molecular Biology, January 2007 doi 10.1590/s1415-47572007000500020

[31] Teruaki Taji, Motoaki Seki, Masakazu Satou, Tetsuya Sakurai, Masatomo Kobayashi,

[32] Kanako Ishiyama, Yoshihiro Narusaka, Mari Narusaka, Jian-Kang Zhu, Kazuo Shinozaki

[33] Comparative Genomics in Salt Tolerance between Arabidopsis and Arabidopsis-Related Halophyte Salt Cress Using Arabidopsis Microarray Plant Physiology Jul 2004, 135 (3) 1697-1709; DOI: 10.1104/pp.104.039909

[34]]Mohammad Aslam,Beenish Fakher ,Bello Hassan Jakada ,Lihua Zhao,Shijiang Cao ,Yan Cheng and Yuan Qin .Genome-Wide Identification and Expression Profiling of CBL-CIPK Gene Family in Pineapple (Ananas comosus) and the Role of AcCBL1 in Abiotic and Biotic Stress Response Biomolecules 2019, 9(7),293; https://doi.org/10.3390/biom9070293

[35] Li L, Xu X, Chen C, Shen Z. Genome-Wide Characterization and Expression Analysis of the Germin-Like Protein Family in Rice and Arabidopsis. *Int J Mol Sci*. 2016;17(10):1622. Published 2016 Sep 23. doi:10.3390/ijms17101622

[36] Sonah, H., Deshmukh, R.K., Labbé, C. *et al.* Analysis of aquaporins in Brassicaceae species reveals high-level of conservation and dynamic role against biotic and abiotic stress in canola. *Sci Rep* 7, 2771 (2017). https://doi.org/10.1038/s41598-017-02877-9

[37] J. M. Zhang, M. Harman, L. Ma and Y. Liu, Machine Learning Testing: Survey, Landscapes and Horizons, *IEEE Transactions on Software Engineering*, doi: 10.1109/TSE.2019.2962027.

[38] Bethany M. Moore, Peipei Wang, Pengxiang Fan, Bryan Leong, Craig A. Schenck,

[39] Robust predictions of specialized metabolism genes through machine learning John P. Lloyd, MelissD. Lehti-Shiu, Robert L. Last, Eran Pichersky, Shin-Han Shiu PNAS March 19, 2019 116 (12) 5830, https://doi.org/10.1073/pnas.1902386116

[40] , H., Zhang, J., Li, J. *et al.* Genome-wide transcriptomic analysis of a desert willow, *Salix psammophila*, reveals the function of hub genes *SpMDP1* and *SpWRKY33* in drought tolerance. *BMC Plant Biol* 19, 356 (2019). https://doi.org/10.1186/s12870-019-1900-1

[41] Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu S-H (2016) Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput Biol 12(12):e1005244. doi:10.1371/journal.pcbi.1005244

[42] Davila Olivas NH, Coolen S, Huang P, Severing E, van Verk MC, Hickman R, Wittenberg AH, de Vos M, Prins M, van Loon JJ, Aarts MG, van Wees SC, Pieterse CM, Dicke M. Effect of prior drought and pathogen stress on Arabidopsis transcriptome changes to caterpillar herbivory. New Phytol. 2016 Jun;210(4):1344-56. doi: 10.1111/nph.13847. Epub 2016 Feb 5

ISSN- 2394-5125 VOL 7, ISSUE 11, 2020

[43] Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change anddispersion for RNA-seq data with DESeq2. Genome Biology 15: 550.

[44] Verma, H., Borah, J.L. & Sarma, R.N. Variability Assessment for Root and Drought Tolerance Traits and Genetic Diversity Analysis of Rice Germplasm using SSR Markers. *Sci Rep* 9, 16513 (2019). https://doi.org/10.1038/s41598-019-52884-1

[45] Razzaq, H., Tahir, M.H., Sadaqat, H.A., & Sadia, B. (2017). Screening of sunflower (Helianthus annus L.) accessions under drought stress conditions, an experimental assay. Int.J.Curr.Microbiol.App.Sci (2017) 6(5): 848-856

[46] Ahmad Tahmasebi a, b, Elham Ashrafi-Dehkordi a, b, Amir Ghaffar Shahriari c, *,

[47] Seyed Mohammad Mazloomi a, Esmaeil Ebrahimie d, e, f, gIntegrative meta-analysis of transcriptomic responses to abiotic stress in cotton Progress in Biophysics and Molecular Biology 146 (2019) 112e122 https://doi.org/10.1016/j.pbiomolbio.2019.02.005

[48] Leek, J.T., Johnson,W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments.

[49] Bioinformatics 28, 882e883.

[50] Kang, D., Ahn, H., Lee, S. *et al.* StressGenePred: a twin prediction model architecture for classifying the stress types of samples and discovering stress-related genes in arabidopsis. *BMC Genomics* 20, 949 (2019). https://doi.org/10.1186/s12864-019-6283-z

[51] Di Zhang, Yi Yue, Yang Zhao, Chao Wang, Xi Cheng, Ying Wu, Guohua Fan,

[52] Panrong Wu, Yujia Gao, Youhua Zhang, Yunzhi Wu. CSTDB: A Crop Stress-tolerance Gene and Protein Database Integrated by Convolutional Neural Networks. bioRxiv; 2018. DOI: 10.1101/456343.

[53] Kim Y., 2014 Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.https://arxiv.org/abs/14