International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

Human Disease Prediction Using Machine Learning Classification Algorithms

Dillip Narayan Sahu[1], Satheesh Kumar[2]
[1]*Research Scholar, Department of Computer Science and Application, OPJS University, Churu, Rajasthan, India*
[2]*Professor, Department of Computer Science and Engineering, Chaitanya (Deemed to be University), Telangana, India*

ABSTRACT
The importance of healthcare system is growing and the pandemic has proved that the healthcare management is an important part of an individual's life. Most medical cases requires proper diagnoses in a prior consultation, and it is very important to get an accurate prediction of the disease. In this paper, we have taken some most common and very serious human diseases as an input in our dataset, after preprocessing, data cleaning, and using different machine learning classification algorithms such as Random Forest, Naïve Bayes, Ada Boost, Bagging Stacking, we found that the Machine learning classification algorithms can be a better option for the early prediction with more accuracy for the human diseases.

Keywords: Algorithm, Classifier, Human Diseases, Machine Learning, Prediction

1. Introduction

Machine learning tools are widely used in all fields of science and medicine and are responsible for revolutionizing businesses everywhere. Healthcare systems, on the other hand, are very slow to adopt these advances and are far behind [1][2]. Machine learning is often useful in the treatment of chronic diseases, namely kidney disease, heart diseases, diabetes etc[3][4]. In fact, machine learning is already being used to predict diabetes risk supported by genomic data, supported by EHR data to diagnose diabetes to predict risk of complications. The introduction of machine learning methods can significantly increase the detection and early treatment of diabetes complications in patients [5][6].
The medical disease prediction application can also be used. basic knowledge about the disease and can tell us if we should seek immediate medical attention for temporary relief or at least start with home remedies. Combining machine learning with an API for user interaction provides an opportunity to facilitate interaction with users by using a machine learning model to make more accurate predictions.[7]
Chronic renal disorder (CKD) is a major burden on the healthcare system because of its increasing prevalence, high risk of progression to end-stage renal disease, and poor morbidity and mortality prognosis. it's rapidly becoming a global health crisis. Unhealthy dietary habits and insufficient water consumption are significant contributors to the present disease. Without kidneys, an individual can only live for 18 days on average, requiring kidney transplantation and dialysis. it's critical to have reliable techniques at predicting CKD in its early stages[8][9]. Machine learning (ML) techniques are excellent in predicting CKD. the present study offers a methodology for predicting CKD status using clinical data, which includes data preprocessing, a way for managing missing values, data aggregation, and have extraction. variety of physiological variables, also as ML techniques such as logistic regression (LR), decision tree (DT) classification, and -nearest neighbor (KNN), were utilized in this work to train three distinct models for reliable prediction. The LR classification method was found to be the foremost accurate in this role, with an accuracy of about 97 percent during this study. The dataset that was utilized in the creation of the technique was the CKD dataset, which was made available to the general public . Compared to prior

DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

research, the accuracy rate of the models employed during this study is considerably greater, implying that they're more trustworthy than the models used in previous studies as well. an outsized number of model comparisons have shown their resilience, and therefore the scheme may be inferred from the study's results[10].

Diabetes is one of the serious diseases and many people suffer from this disease. Aging, obesity, lack of exercise, hereditary diabetes, lifestyle habits, unbalanced diet, hypertension, etc., can cause diabetes. People with diabetes are at high risk for diseases such as heart disease, kidney disease, stroke, eye disease, and nerve damage. Current practice in hospitals is to collect the information necessary for diagnosing diabetes through various tests, and based on that, the diagnosis and appropriate treatment are made. Big data analytics plays an important role in the healthcare industry. The healthcare industry has a huge database. Big data analytics can be used to explore massive data sets, find hidden information and patterns, discover knowledge from data, and predict corresponding outcomes. Existing methods are not very accurate in classification and prediction. In this article, we proposed a diabetes prediction model to better classify diabetes. It contains few external factors that contribute to diabetes besides the usual factors such as glucose, BMI, age and insulin. The new dataset has better classification accuracy compared to the existing dataset. Additionally, a diabetes prediction pipeline model was applied to improve classification accuracy.

2. Experiments and Observations

We have used different dataset for different diseases which are in csv as well as arff format, for the analysis, we have taken Weka machine learning tool and also for the purpose of preprocessing, cleaning, classification and accuracy acceptance comparative analysis purpose.

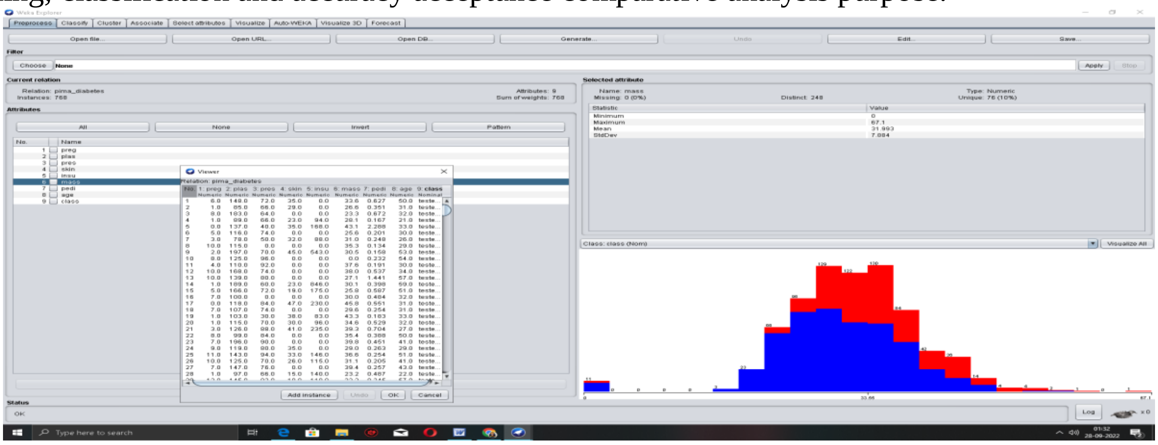

Fig.1 Preprocess of Diabetes Dataset having 10 Attributes

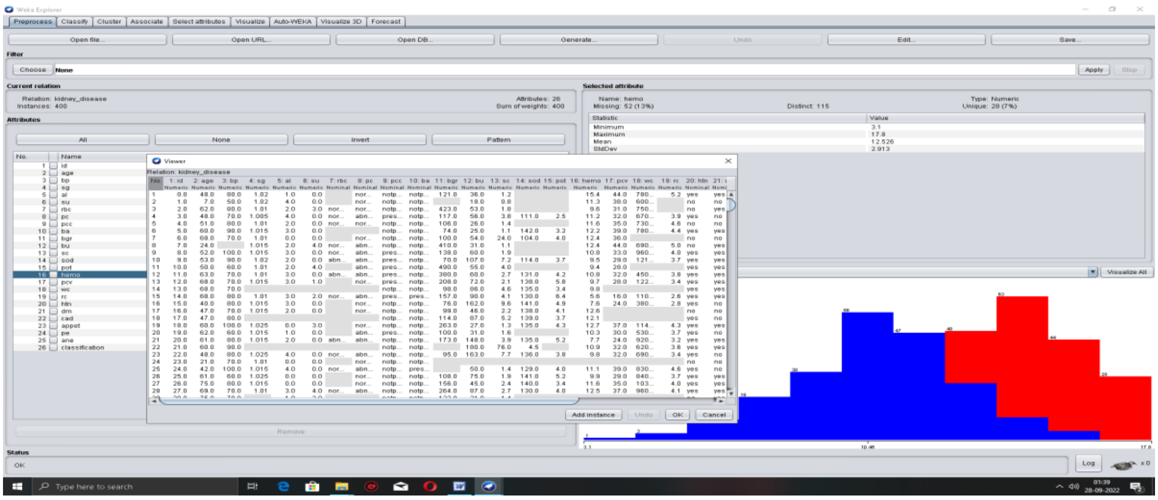

Fig.2 Preprocess of Kidney Dataset having 26 Attributes

Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

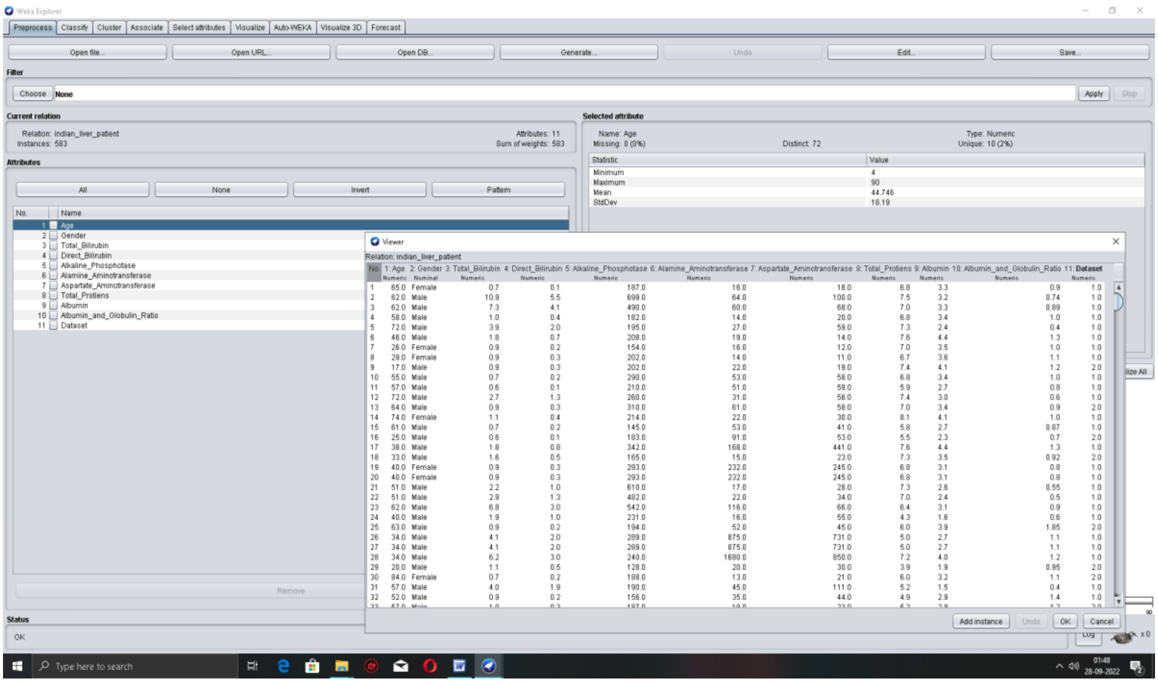

Fig.3 Preprocess of Liver Dataset having 11 Attributes

Algorithm taken- UltraBoost
Classifier Output-=== Run information ===
Scheme: weka.classifiers.meta.UltraBoost -S 1 -B "weka.classifiers.meta.FilteredClassifier -F \"weka.filters.unsupervised.attribute.RemoveType -V -T nominal\" -S 1 -W weka.classifiers.bayes.NaiveBayes" -B "weka.classifiers.meta.FilteredClassifier -F \"weka.filters.unsupervised.attribute.RemoveType -V -T numeric\" -S 1 -W weka.classifiers.functions.Logistic -- -R 1.0E-8 -M -1 -num-decimal-places 4"
Relation: pima_diabetes
Instances: 768
Attributes: 9
preg plas pres skin insu mass
pedi age class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
UltraBoost
Base classifiers
FilteredClassifier using weka.classifiers.bayes.NaiveBayes on data filtered through weka.filters.unsupervised.attribute.RemoveType -V -T nominal
Filtered Header
@relation pima_diabetes-weka.filters.unsupervised.attribute.RemoveType-V-Tnominal
@attribute class {tested_negative,tested_positive}
@data
Classifier Model
Naive Bayes Classifier
Class
Attribute tested_negative tested_positive
(0.65) (0.35)
===
FilteredClassifier using weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4 on data filtered through weka.filters.unsupervised.attribute.RemoveType -V -T numeric

```
Filtered Header
Classifier Model
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                    Class
Variable     tested_negative
=============================
preg              -0.1234
plas              -0.0351
pres               0.0131
skin              -0.0004
insu               0.0012
mass              -0.0901
pedi              -0.9771
age               -0.0159
Intercept          8.2873

Odds Ratios...
                    Class
Variable     tested_negative
=============================
preg               0.8839
plas               0.9656
pres               1.0132
skin               0.9996
insu               1.0012
mass               0.9138
pedi               0.3764
age                0.9842

Time taken to build model: 0.03 seconds
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances         590               76.8229 %
Incorrectly Classified Instances       178               23.1771 %
Kappa statistic                          0.4458
Mean absolute error                      0.3795
Root mean squared error                  0.4124
Relative absolute error                 83.4983 %
Root relative squared error             86.5164 %
Total Number of Instances              768
=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC     ROC Area  PRC Area  Class
               0.914    0.504    0.772      0.914    0.837      0.465   0.832     0.892     tested_negative
               0.496    0.086    0.756      0.496    0.599      0.465   0.832     0.715     tested_positive
Weighted Avg.  0.768    0.358    0.766      0.768    0.754      0.465   0.832     0.830
=== Confusion Matrix ===
   a   b   <-- classified as
 457  43 |   a = tested_negative
 135 133 |   b = tested_positive
```

DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

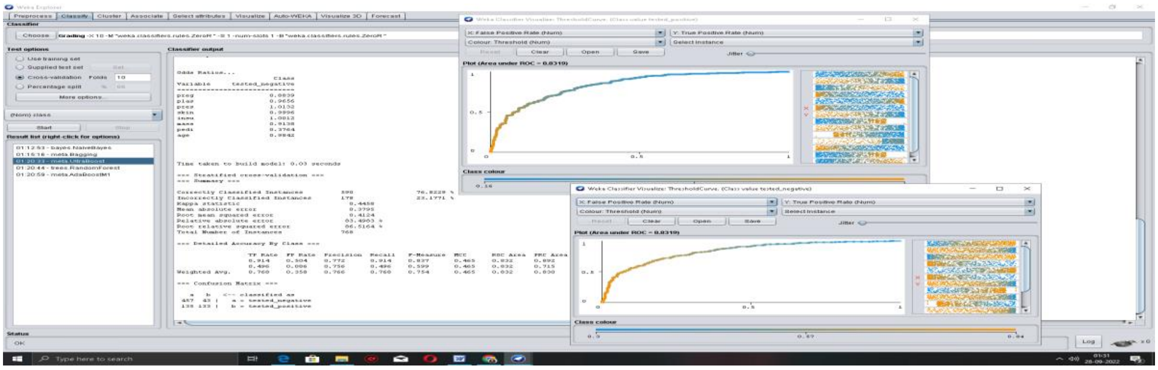

Fig.4 UltraBoost Classifier with Visualize curve (for Diabetes disease)

Algorithm taken- AdaBoostM1 (for Kidney Disease)
Classifier Output=== Run information ===
Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump
Relation: kidney_disease
Instances: 400
Attributes: 26
id age bp sg al su rbc pc
pcc ba bgr bu sc sod pot
hemo pcv wc rc htn dm cad
appet pe ane
classification
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
AdaBoostM1: No boosting possible, one classifier used!
Decision Stump
Classifications
id <= 249.5 : ckd
id > 249.5 : notckd
id is missing : ckd
Class distributions
id <= 249.5
ckd notckd
1.0 0.0
id > 249.5
ckd notckd
0.0 1.0
id is missing
ckd notckd
0.625 0.375
Time taken to build model: 0 seconds
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances 398 **99.5 %**
Incorrectly Classified Instances 2 0.5 %
Kappa statistic 0.9893
Mean absolute error 0.005
Root mean squared error 0.0707
Relative absolute error 1.0663 %

DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

Root relative squared error 14.6059 %
Total Number of Instances 400
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.996	0.007	0.996	0.996	0.996	0.989	0.995	0.995	ckd
	0.993	0.004	0.993	0.993	0.993	0.989	0.995	0.989	notckd
Weighted Avg.	0.995	0.006	0.995	0.995	0.995	0.989	0.995	0.993	

=== Confusion Matrix ===
a b <-- classified as
249 1 | a = ckd
1 149 | b = notckd

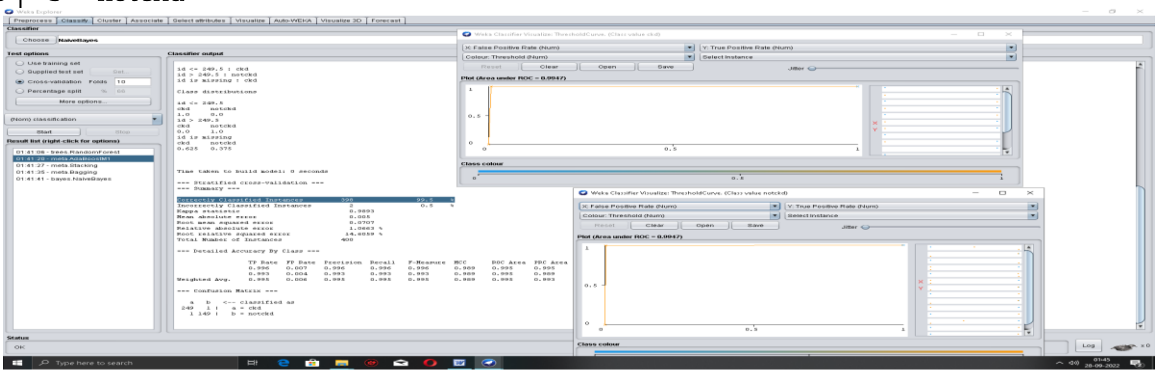
Fig.5 AdaBoostM1 Classifier with Visualize curve (for Kidney disease)

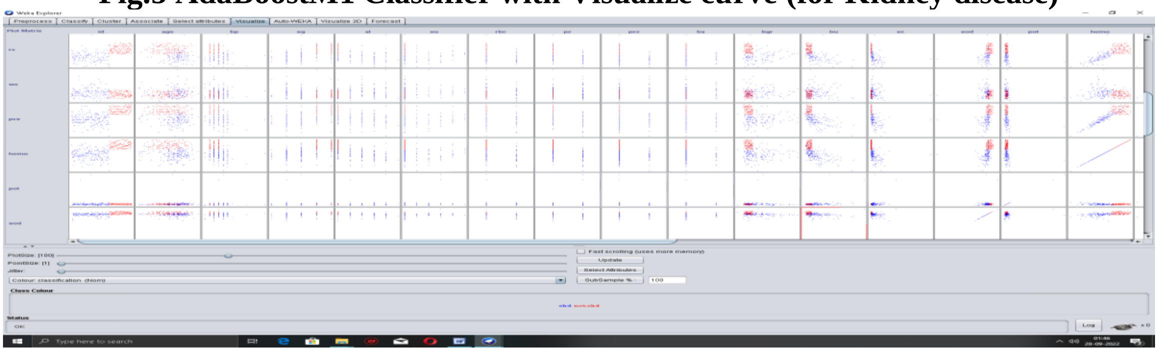
Fig.6 Visualize curve with all the attributes (for Kidney disease)

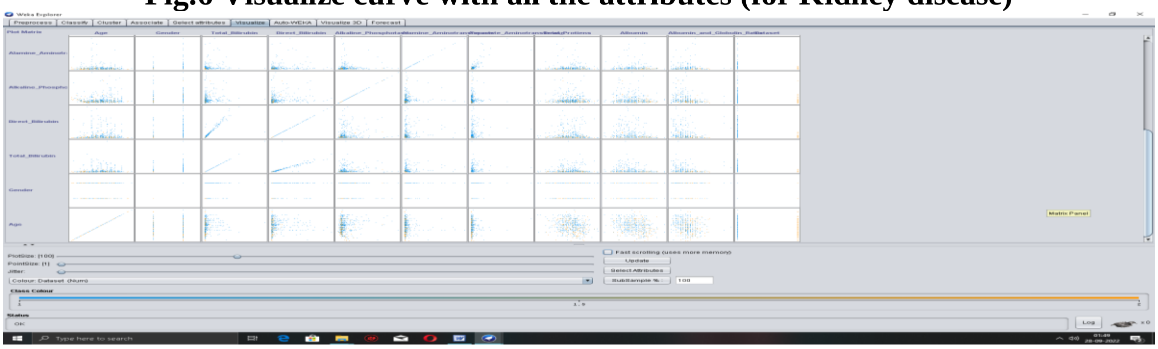
Fig.7 Visualize curve with all the attributes (for Liver disease)

3. Discussion

We have taken 5 different machine learning algorithms and with the help of different experimental observations using the machine learning tool, it is found clearly that the ML is giving good accuracy rate to analyze, detection and prediction for the different human diseases. In the above observations, it is seen that, machine learning algorithms are no doubt an excellent method to predict different human diseases at an early stage. It is found that the accuracy level is acceptable and so will be efficient for the medical sciences.

DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

4. Conclusion

In the study of the above real time medical dataset analysis, experimentation and observation in different parameters of algorithms, it is found that the accuracy level using the machine learning classification model AdaBoostM1 is much satisfactory, having good accuracy rate of 99.5% (for kidney disease) and so will be a good option in the field of medical health care sector to opt or to predict early prediction with proper diagnosis of different human diseases.

References

1. Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B. & Yang, C. W. (2013). Chronic kidney disease: global dimension and perspectives. The Lancet, 382(9888), 260-272.
2. Ali, S., Dave, N., Virani, S. S., & Navaneethan, S. D. (2019). Primary and secondary prevention of cardiovascular disease in patients with chronic kidney disease. Current Atherosclerosis Reports, 21(9), 1-9.
3. Levey, A. S., Coresh, J., Bolton, K., Culleton, B., Harvey, K. S., Ikizler, T. A. & Briggs, J. (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. American Journal of Kidney Diseases, 39(2 SUPPL. 1), i-ii+.
4. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
5. Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 291-296). IEEE.
6. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
7. Noble, W. S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12), 1565-1567.
8. Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics, 9(4), 705-724.
9. Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. International Journal of Advances in Engineering & Technology, 7(1), 242.
10. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2), 1-12.
11. Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. International Journal of Engineering Research and Technology, 4(7), 206–210.