Analysis on Counterfeit Note Detection Using Binary Classification Techniques

Gaythri Mol Shajimon
Computer Science and Electronics Engineering
University Of Essex
Colchester, United Kingdom
gi22846@essex.ac.uk

***Abstract*—Despite of increase in cashless transactions, bank currency still plays a vital role in the financial market. Counterfeit notes are a major threat to transactions using currency. Using advanced technologies it is not at all difficult nowadays to make counterfeit notes with high precision and resemblance with original notes and in most of the cases, identifying these using human-eye is challenging. Bank note authentication machines are available now, but public has limited access to it. So it difficult to stop the circulation of fake notes unless these technologies were available to all.**

Implementation of cost-effective and trusted web or mobile applications and making them available to the public can reduce this circulation of fake notes to a great extent.

Supervised machine learning algorithms can be used to evaluate the authenticity of bank notes. Binary classification is one such supervised learning method used for predicting possible classes of instances into two, based on the classification rules.

Decision tree, KNN, Logistic regression, Random forest classifier, Support Vector Classifier, Naive Bias, and XG boost are some of the classification models that are used in this paper for finding the best suitable model for making bank note authentication system.

***Impact Statement-* SVM performed very well with 100% accuracy score with zero precision and recall and their remaining evaluation matrices were best compared to other models.**

***Index Terms*—Binary Classification, Exploratory Analysis, Evaluation Matrices, Classification report, Confusion Matrix**

I. INTRODUCTION

BAnk currency is the most important asset of a country. Even though most of our transactions are now converted to cashless, still currency transaction plays a major role in the global market. Miscreants make counterfeit notes with high precision and resemblance to the original currencies and will cause discrepancies and imbalances in the financial market. Despite of having lots of available advanced technologies, still there is a drastic increase in the circulation of fake notes in the market.

As fake notes are majorly used for illegal activities, it is a matter of national security as well. So it is important to introduce security features to mitigate counterfeit notes. There are a lot of research works currently in progress to safely implement security features in currencies, that are not externally visible so it will be difficult to copy in the forged note.

For smooth cash transactions and to determine the legitimacy of the currency, a cost-effective counterfeit note authentication system should be implemented and made available to pubic use. In this way, we can prevent this unlawful activity to a great extent.

When conventional mathematical methods fail to solve this problem, artificial intelligence and machine learning algorithms can play a crucial role in mitigating this activity. We can make use of supervised learning techniques in this case. Due to the nature of the problem, binary classification techniques will be suitable for this type of classification. Using image processing techniques the application should be able to extract the features and using these features as input the model should be able to predict the fake or original note. Binary classification techniques such as Decision Tree, SVM, KNN, Random Forest Classifier, XG Boost, Naive Bayes, and Logistic Regression are some of the classification techniques that we can use for this problem. The best binary classifying technique will be identified at the end using the model performance, evaluation matrices, and reports.

The remaining section in this paper is organized as follows. Section 2 contains the literature survey, which gives an idea of recent developments and current researches in bank note authentication systems and also gives an idea of different methodologies that can be used for implementing the system. Section 3 deals with the Methodology, which describes the methods that followed in the project including prepossessing, training and evaluation of model. Section 4 is Result, which give the final result and observation got at the end of analysis. Section deals with the discussion, which contains the lessons learned after completion of project, challenges faced and limitations. Final section is Conclusion were the whole work done and its in the project is summarized.

II. LITERATURE REVIEW

Safeguarding the authenticity of higher denominations of printed currencies is one of the critical issues nowadays. Currencies have a vital role in the financial activities of every country [1]. Exhaustive studies has been done for the evaluation and concludes that Decision-Tree and MLP were the best classification techniques to classify a banknote. In an another study [2], features of currencies were extracted using method called Fast Wavelet Transformation. It classifies currencies in to four different types: High Quality Forgery, Genuine, Low Quality Forgery and Inappropriate ROI, and it has a results of 100% accuracy. K means clustering was used in [3], and suggest that, the data could be better processed, if we have more features, means with less features probability of precision and recall will be more. Smart phones were use for bank note authentication in [4]. It is based on constructive adaptive wavelet for analysing different print patter in bank notes and generated a linear and stable output.

Image processing and pattern recognition can be used to design authentication machine [5]. The importance of embedded security aspects were also analysed as part of research and got high accuracy score and they recommended government to introduce the new embedded security features to reduce counterfeit notes. Support vector Machine and Back propagation techniques were used in [6]. An extensive research has been done to find the best model and concluded that SVM performs well compared to back propagation. A different type of research has been done in [7] using touch, vision and both touch and vision with limited time. They have chose experts and non- experts for the study. The experiment with either touch or vision results in poor result compared to both vision and touch, especially if longer duration is allowed. A comparative study has been done with SVM, Naive Bayes and ANN in [8] for identifying counterfeit note using same UCI data. They have used hold out and cross validation methods, result was Multi layer perception outperform Naive Bayes in term of accuracy score, but Naive Bayes is faster compared to Multi Layer Perception.

As most of currently available authentication systems hardware related and costly it is difficult make use of it every one [9]. Using security threads, latent images and water marks a study conducted to authenticating counterfeit note. Models used for training are KNN, SVM, Random Forest and Logistic regression. SVM outperform other models in terms of accuracy. There is a lot of research happening in this area, especially in implementing the security features in bank note. Security feature that are adding must make a bank note unique and won't be able to duplicate.

III. Methodology

UCI Machine Learning Repository is the source of the data set used to train the models. Data were extracted from original and fake banknote images. The data set has 1372 rows and 5 columns, out of which 4 are features and one is target class. The ratio between the original and counterfeit note is 56:44 (original: fake), The target class contains values 0 and 1, where 0 represents the original note and 1 represents the counterfeit note. Detailed description of data set is described in Table 1.

Hold out method is used to train the model, where the whole data set is divided into train, test data set.

Data set was slightly imbalanced, so in order to get the accurate prediction data set has been balanced. Exploratory analysis has been done for finding null values, outlier detection, linearity, co-relation etc along with graphical representation. Validation set is not used, as grid search and cross validation is used for tuning and generalisation. In cross validation it internally will divide training data in to k folds of training and validation set, so it is not relevant to divide data set in to validation set and for hyper-parameter tuning grid search. Models were trained using train data set with best parameter given by grid search. Cross validation help to identify the actual performance of model and helped to find the best model. Training has been done for 7 classification models including Decision Tree, SVM, KNN, Naive Bayes, Logistic Regression,Random Forest and XG Boost. The evaluation matrices used to find the best model is Cross validation score, Accuracy score, F1 Score, F Beta Score, Precision Score, Accuracy Score, ROC AUC Score. The final model evaluation is done using Test data set, which 20% of total data set. The models that were considered for training are described below:

i. Decision Tree

A Decision Tree is a supervised learning algorithm where instance space is partitioned recursively. Unlike other supervised learning algorithms it can be used for solving both classification and regression problems. A decision tree is a directed tree with roots and nodes with the parent node is called a root, and it has no incoming edges and all other nodes will have exactly one incoming edge. Nodes with no child node or outgoing edges are called leaf nodes (Terminal node) and nodes with outgoing edges are called test nodes (Internal node). According to certain discrete functions, we can divide each test node into sub-spaces of input attribute values [1]. The most appropriate target values of a class will be assigned to a leaf. We can classify instances by navigating them from the root of the tree to its leaf. In this way using a Decision tree we can train a model to predict the class of target attribute by learning decision rules from training data(prior data)

Table 1 . Data set Description

Attribute	**Type**	**Description**
Variance of Wavelet Transformed image	continuous	Variance is the average value of measure of spread of a distribution
Skewness of Wavelet Transformed image	continuous	It is a measure of lack of symmetry
Kurtosis of Wavelet Transformed image	continuous	Kurtosis is described by the peakedness of distribution
Entropy of Wavelet Transformed image	continuous	Image entropy uses compression algorithm which describes the amount of information which must be coded.
Output of class	Integer	Output contains two values, 0 representing Original notes and 1 representing counterfeit note

ii. Support Vector Machine

A support Vector Machine comes under supervised learning algorithm were it is used to evaluate the data and recognize the patterns to which data needs to be classified. A decision boundary will be created to evaluate the model. In the Support Vector Machine, each instance will be plotted on a graph and differentiate into two classes by a hyperplane, using classification.

In order to make non-linearly separable data into linearly separable data we can make use of Kernel functions in SVM, where we will project the data from lower-dimensional space to higher-dimensional space. As there are only a few features and more number of test cases, linear kernel isused for classification. As the data set is linearly separable, a linear kernel can find a linear margin that can separate the graph into two regions.

The decision boundary is chosen in such a way that it has a maximum distance from data point [6].

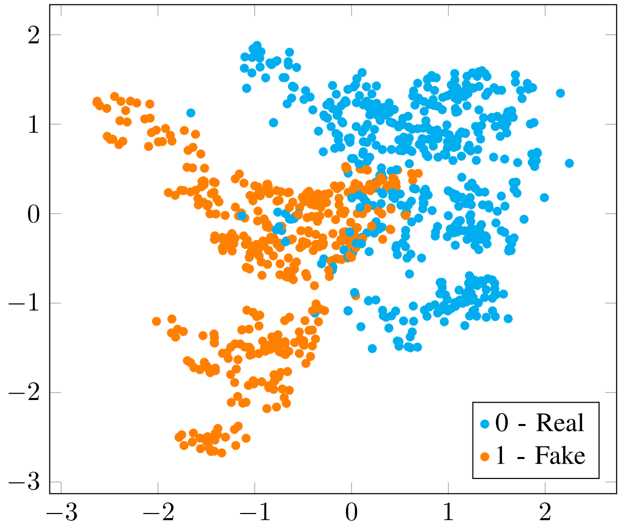

Figure 1 . SVM Visualisation

iii. Random Forest Classifier

Random forests are a group of a large number of decision trees where each tree depends on the random vector values which are sampled independently and have the same distribution for all trees in the forest. It can efficiently handle large data sets. Various decision tress will be trained using the training data. And data set contains randomly chosen features and observations selected during the splitting of nodes. Leaf node of each decision tree is output produced by that decision tree. In this way output produced by each decision tree will be used for the final result. A majority voting system will be used here. The output given by majority of the decision trees will be considered as final predicted result.

iv. Naive Bayes

A Naive Bayes classifier is used when all features in the data set are independent of each other. Even though Independence is not a good assumption, Naive Bayes performs well in actual practice [10]. For binary classification When the number of target classes goes beyond two, Naive Bayes will be the best option and that makes its position more relevant. Some of the successful real-life application of Naive Bayes includes whether prediction, customer credit evaluation, etc.

Naive Bayes is a mathematical classifier, which can classify the data set only if the data set of the problem is prepossessed in tabular format. In this way, the best-fitted classification will be the one with the highest fitness value [11].

v. K Nearest Neighbor

K Nearest neighbor algorithm(KNN) is a most fundamental and simple classification algorithm mainly used when there is only little or no prior knowledge about the distribution of data. It is developed from the need to do discriminant analysis when reliable parametric estimates of probability densities are not reliable [12]. KNN classifies the data point by checking how its neighbors are classified. It uses the similarity measures of the earlier stored data points for classification. It will store the features of the current data set and use this for comparison and classifies when new data came for classification.

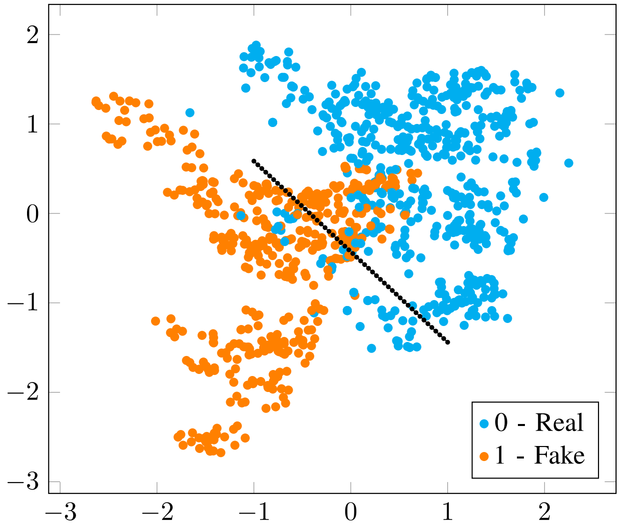

Figure 2 . Decision Boundary - SVM

vi. Logistic Regression

Logistic Regression is also a supervised learning algorithm in which, it will predict the classes using dependant variables. Logistic Regression is a blend of two types of statistical tradition. One is the analysis of crosstabs and the other in which all variables are measured at either the two categories, nominal or ordinal in which variables usually have few distinct categories. In logistic regression, we use a sigmoid function to transform the output value of the algorithm into probability value and the target class is predicted using probability value. The hypothesis of Logistic regression states that the cost function should be restricted between the values 0 and 1.

vii. XG Boost

XG Boost is the short form of Extreme Gradient Boosting. it uses Gradient boosting frame work. XG Boost supports various functions like ranking, classification and regression. The package is flexible to can be extended, such that we can create our own objectives easily. It is supported by almost all major programming languages.

XG Boost is highly efficient and flexible machine learning algorithm and also it is scalable and distributed GBDT (Gradient-boosted decision tree) and can solve data science problems quickly and accurately compared to other models. In XG Boost parallel tree boosting is happening and that makes it leading library for classification problems.

XG Boost is also a gradient-boosting algorithm, the only difference is unlike in gradient boosting algorithms, weak learner addition processes does not happen one after the other, instead, a multi-threaded approach will be used and proper utilization of the CPU core leads to greater performance and speed.

Some of the limitations of XG Boost is it won’t perform well on unstructured data and on sparse and it is very sensitive towards outliers. So chances of getting error will be more if outliers are present in the data. One of the other limitation is, it s not easily scalable.

Table 2. Evaluation Matrix & CV Scores in %

Evaluation Matrix and Cross-Validation scores				
Models	**Accuracy**	**Precision**	**Recall**	**CV**
Decision Tree	97.81	97.43	97.43	98.63
SVM	100	100	100	100
Random Forest	98.18	96.20	99.14	99.45
KNN	100	100	100	99.73
Naive Bayes	83.63	83.33	76.92	84.49
Logistic Regression	100	100	100	99.73
XG Boost	90.09	89.65	88.88	99.18

Table 3. Confusion Matrix Results

Confusion Matrix Results				
Models	**TP**	**TN**	**FP**	**FN**
Decision Tree	144	155	0	4
SVM	158	117	0	0
Random Forest	154	116	1	3
KNN	158	117	0	0
Naive Bayes	140	90	27	19
Logistic Regression	158	117	0	0
XG Boost	146	89	13	12

IV. Result

Extensive study has been done to find model for implementing bank note authentication system. Among 7 models, Support Vector Machine, K Nearest Neighbor and Logistic Regression ends with 100% accuracy score, precision, recall, f1 score, f beta score, roc-auc score. But surprisingly in cross validation, when comparing the cross validation score, only SVM has 100% accuracy among all other models. Naive Bayes classifier got least score in all the evaluation matrices compared to others. Also noticed that Random Forest Classifier, is very slow compared to other model.

Table 2. depicts the different metrics that have been used for evaluating the model. Accuracy, precision, Recall, f1 score, roc-auc score has been calculated for all the models.

Table 3. depicts the number of true positive, true negative, false positive, false negative values obtained using the confusion matrix. It gives us a better visibility of performance of model. Considering confusion matrix values SVM, KNN, and Logistic Regression predicted all real and fake with 100% accuracy.

Figure 3. depicts the graph plotted with roc-auc scores of each model. From this graph we can see that SVM, RFC, KNN, LR has roc-auc score of 100% and Naive Bayes has least score among all.

KNN, SVM, and Logistic Regression performed equally with 100% accuracy in all evaluation matrix. As we have to find one best model among all, considering cross validation score as well, so came to conclusion that Support Vector Machine classifier is the best model among all, with KNN and Logistic Regression in second position with slight difference in cross validation score.

Figure 1. and Figure 2. is the scatter plot diagram of SVM with orange color represents fake notes and cyan color represents real notes. In figure 2. we can see a black color line, which is decision boundary used to separate real and fake notes.

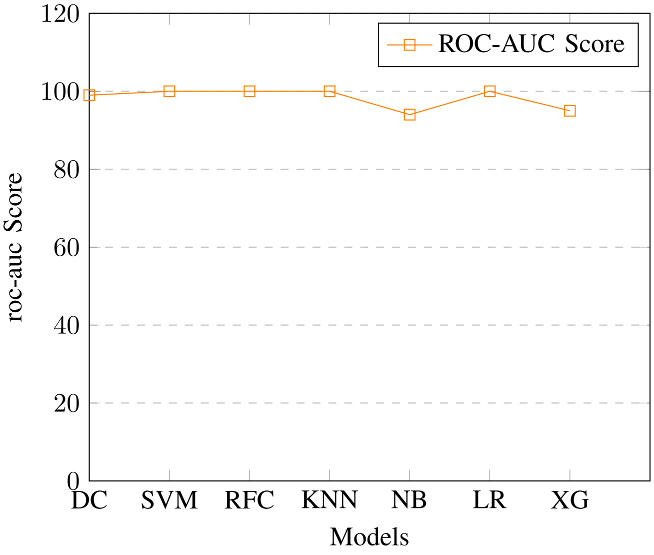

Figure 3. ROC-AUC score

V. Discussion

For learning, the whole data set is divided into 80% training and 20% test set. Out of 275 data in test data 158 were fake and 117 were original.

Using image processing techniques and feature extraction 4 most important features of bank notes were extracted and was given as input to the model for training. The system evaluated all the models with the given 4 features and predicted the output. All classifier except Naive Bayes and XG Boost gave more than 98% accuracy and performed well in the training. Figure 3. depicts the confusion matrix result of SVM classifier. Out of 7 models trained, logistic regression, K Nearest Neighbor gave 100% scores but when comparing cross validation score SVM only has 100% cross validation score among all. Random Forest classifier and Decision Tree classifier performed well but only ends with accuracy of 98%.

Lots of research works are in progress to find the best technique for classifying the real and fake notes. Even though lot of techniques are already available, they were mostly hardware related, common people have no access to most of them. Web and mobile applications are available , but most of them are not 100% accurate and trusted. Even though some applications performed well, but it is difficult to get 100% accuracy for all denominations.

Processing the image of each bank note and doing analysis to find the counterfeit note is practically difficult and time consuming. As there are many advanced techniques to mock original notes even with high precision and high resemblance to original currencies, security features should be introduced in the bank notes like RFID, electronic chips etc, so that rather than image processing we can make use of these security features for finding the counterfeit notes. Lots of researches were going on in this area currently. The main bottleneck for introducing these feature is the cost. Demonetization of currently using using bank notes might be required if we have replace the currently using currencies and cost of making each banknote with above mentioned security features will be an extra cost for the country. So this can be implemented in slow pace only. Till that time in order to prevent the circulation of counterfeit note, we need to rely on image processing techniques.

VI. Conclusion

With increase in challenges raised by counterfeit notes, implementing best solution for it has great relevance. For identifying the best model for implementing bank note authentication system, seven models were trained using 80% of the data, and evaluated the trained model using 20% of the test data. Among all seven model, Support Vector Machine is selected as the best model with 100% accuracy in all scores and evaluation matrices. Even though, Logistic Regression and KNN has same score of 100%, cross validation score is less than SVM for others. Random forest and Decision tree also performed well but ends with an accuracy of 98%. Models which gave least accurate predictions were Naive Bayes and XG Boost with accuracy score of 84% and 90%. So we can conclude that best model among all seven model is Support Vector Machine Classifier.

References

[1] C. Kumar and A. K. Dudyala, "Bank note authentication using decision tree rules and machine learning techniques," in *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 310–314, IEEE, 2015.

[2] E. Gillich and V. Lohweg, "Banknote authentication," *1. Jahreskolloquium Bild. Der Autom*, pp. 1–8, 2010.

[3] E. Ragavi, "Banknote authentication analysis using python k-means clustering," *International Journal of Innovative Science and Research Technology*, vol. 5, no. 10, pp. 80–82, 2020.

[4] V. Lohweg, J. L. Hoffmann, H. Dörksen, R. Hildebrand, E. Gillich, J. Hofmann, and J. Schaede, "Banknote authentication with mobile devices," in *Media Watermarking, Security, and Forensics 2013*, vol. 8665, pp. 47–60, SPIE, 2013.

[5] A. Roy, B. Halder, U. Garain, and D. S. Doermann, "Machine-assisted authentication of paper currency: an experiment on indian banknotes," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 3, pp. 271–285, 2015.

[6] S. Shahani, A. Jagiasi, and R. Priya, "Analysis of banknote authentication system using machine learning techniques," *International Journal of Computer Applications*, vol. 975, p. 8887, 2018.

[7] F. van der Horst, J. Snell, and J. Theeuwes, "Finding counterfeited banknotes: The roles of vision and touch," *Cognitive Research: Principles and Implications*, vol. 5, no. 1, pp. 1–14, 2020.

[8] A. Ghazvini, J. Awwalu, and A. A. Bakar, "Comparative analysis of algorithms in supervised classification: A case study of bank notes dataset," *International Journal of Computer Trends and Technology*, vol. 17, no. 1, pp. 39–43, 2014.

[9] S. Gopane and R. Kotecha, "Indian counterfeit banknote detection using support vector machine," in *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*, 2020.

[10] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.

[11] K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, no. 60, pp. 1–8, 2006.

[12] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.