Comparative Analysis Of Machine Learning Based Bank Note Authentication Through Variable Selection

Rick Nie

Received June 29, 2023
Accepted November 05, 2023
Electronic access December 15, 2023

Banknote authentication plays a crucial role in maintaining the integrity of financial systems. Counterfeit and genuine banknotes are almost identical and nearly impossible to discern from the human eye. With the advancements in machine learning and the availability of diverse variables, it becomes essential to identify the optimal combination of variables and machine learning algorithms for accurate banknote authentication. This research paper employs the UCI machine learning dataset, which contains properties of images of genuine and counterfeit banknotes. A systematic analysis was conducted to identify the most informative subsets of variables for authentication. After rigorous preprocessing, including data cleaning, normalization, and feature scaling, Six machine learning algorithms were compared. The results revealed that K-Nearest-Neighbors and Support Vector Machine with Gaussian Kernel are the most accurate algorithms for banknote authentication. This study underscores the significance of selecting an appropriate combination of variables and algorithms based on the dataset and suggests that these algorithms can be integrated into mobile applications, especially in developing countries where traditional banking infrastructure is limited. This would provide a cost-effective solution to combat the evolving challenges of counterfeiting.

Introduction

Traditional methods of detecting forgery in banknotes rely on manual inspection and various security features designed into genuine banknotes. With counterfeiters constantly evolving their techniques to create fake banknotes that closely resemble genuine ones, it becomes more and more challenging for traditional methods to keep up as these counterfeit banknotes may include almost the same intricate patterns, colors, and security features as the real ones.

Numerous research studies have been conducted to develop automated systems that can detect counterfeit currency to combat forgery money. Often machine learning-based, these systems employ various techniques such as image analysis, pattern recognition, and classification algorithms to identify counterfeit banknotes based on their visual features, security features, or a combination of both[1,2].

These machine learning systems' successes depend on the data quality and the learning algorithms' performance. The datasets, whether generated by the researcher themselves or readily available by downloading from public domain, were all created by taking pictures of both real and fake banknotes and using image processing techniques and feather extraction tools. As different countries have different security features for their banknotes, researchers must generate their data from the specific banknotes that they are interested to authenticate, and algorithm, such as Convolutional Neural Networks (CNN)[3], Support Vector Machine (SVM)[4] or fuzzy logic[5] were then used to detect the fake currency. Comparing to generating your own dataset, which is time and resource consuming, readily available data set, such as UCI machine learning repository[6] is a good option if the researchers are interested in how different algorithms affect the outcome of banknote authentication[3,7,8].

It is known that both the dataset and algorithms are critical to the success of the machine learning model. However, most of the researchers using the UCI dataset focused mostly on comparing the performances of different algorithms, and few researchers explored the impact of the features in the UCI dataset on the performances of specific algorithms .In this article, various machine learning techniques will be applied to create models which utilize different algorithms and combinations of distinctive banknotes features.

The Data Set

The dataset used to train the models is taken from the UCI machine learning repository. Its data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images. The dataset has 1372 instances and five attributes. Among the five variables, four are features, and one is target class. These five variables are described in Table 1 below. The

four features, Variance, Skewness, Curtosis, and Entropy, are continuous numbers that measure the characteristics of digital images of each banknote. The target class contains two values, 0 and 1, where 0 represents a genuine note, and 1 represents a fake note. The dataset contains a balanced ratio of both classes which is 55:45 (genuine: counterfeit).

variance	skewness	curtosis	entropy	class
3.6216	8.6661	-2.8073	-0.44699	0
4.5459	8.1674	-2.4586	-1.4621	0
3.866	-2.6383	1.9242	0.10645	0
3.4566	9.5228	-4.0112	-3.5944	0
0.32924	-4.4552	4.5718	-0.9888	0
4.3684	9.6718	-3.9606	-3.1625	0

Fig. 1 First 6 rows of the data

Methodology

Several prepossessing steps were applied to the data set, including data cleaning, normalization, and feature scaling. These preprocessing steps are needed to prepare the data set for machine learning algorithms by ensuring the data format is in a suitable range for the algorithms.

Data cleaning was applied to eliminate any possible inconsistent data entries that can negatively affect the performance of the machine learning algorithms. In the data cleaning process, the data set was examined and checked for any missing and potentially erroneous or inconsistent data to ensure the data was clean and complete and to avoid bias or inaccuracies in the subsequent analysis.

The input features were normalized by transforming the range of each axis into a unit scale and removing the mean, as the input variables all have different scales. The normalization process brings all variables to a similar scale, preventing certain variables from dominating others in later analysis and ensuring fair comparisons and interpretations. This process is essential when distance-based algorithms are sensitive to the features' magnitude when using K-Nearest Neighbor or gradient-based algorithms such as logistic Regression.

Scaling the features ensures that each variable is treated fairly by the machine learning algorithm. For example, consider a distance-based algorithm (KNN) with a Euclidean distance function $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. Two features A, B which have magnitude of different scales, notice when plugged into the distance function $\sqrt{(A_x - A_y)^2 + (B_x - B_y)^2}$. The variable with the larger magnitude will dominate the distance function even if it does not contribute significantly to the classification. Similar problems can be observed with gradient-based algorithms (Linear Regression). Therefore, feature scaling ensures that both features A, and B contribute equally to the distance. This paper applied the StandardScaler scaling algorithm to each attribute: variance, Skewness, curtosis, and entropy. It is used because the features are shown to follow a Gaussian distribution and are in different units and scales.

Correlation analysis was applied to each variable to identify the most relevant attributes contributing significantly to the discrimination between genuine and counterfeit bank notes. The data set was split into training and testing sets using a ratio of 80:20, respectively. The training set was used to train the models, while the testing set was used to evaluate the Model's generalization capabilities.

Six machine learning algorithms [13], Logistic Regression, Linear Discrimination Analysis, K- nearest neighbors, Support Vector Machine (Linear Kernel), Support Vector Machine (Polynomial Kernel), and Gaussian Naive Bayes were compared.

Logistic Regression

Logistic Regression predicts the probability of an event or the likelihood of an observation belonging to a particular class. The outcome variable in logistic Regression is binary, usually represented as 0 or 1, where 0 represents the negative or non-event outcome, and 1 represents the positive or event outcome. It is used for binary classification problems, where the goal is to predict a binary outcome or assign an observation to one of two classes. The output of logistic Regression is a probability score for the category prediction, and by providing a probability score for each note's authenticity, logistic Regression provides a more nuanced view rather than a strict classification, which can be helpful in scenarios where a human judgment is applied.

Linear Discrimination Analysis (LDA)

LDA is a feature extraction and dimension-reduction technique that can be used for classification tasks. LDA assumes the class separation is linear and the data is normally distributed. LDA works by maximizing the separation between two classes while minimizing the in-class variation. LDA was chosen because banknotes are manufactured to be consistent in size, features, and security elements. This means banknotes of each denomination and series should be similar. Due to such standardization and precision, banknote features, when plotted, should tend to cluster closely together, while counterfeit banknotes, which cannot replicate the features with high accuracy, will separate from the cluster.

K-Nearest Neighbors (KNN)

KNN, unlike other algorithms used here, is not model-based and does not build an explicit probabilistic or mathematical model from the training data. Rather, KNN calculates distances

Table 1 Description of attributes

Attribute Name	Value Type	Description
Variance	Continuous	Measures how much each pixel varies from the neighboring pixels and classifies them into different regions [9]
Skewness	Continuous	Measures the lack of symmetry [10]
Curtosis	Continuous	Measures whether the data are heavy-tailed or light-tailed relative to a normal distribution [10]
Entropy	Continuous	Describes the amount of information that must be coded for by a compression algorithm [11]
Class	Integer	Consists of two values, 0 representing genuine note and 1 representing fake note [12]

between the new instance and existing instances in the training set to identify the k-nearest neighbors. The distance metric used measures the similarity between instances in the feature space. Determining the K-nearest neighbors also identifies the most similar known instances in terms of feature values. Then, KNN uses the most frequently occurred class labels to assign the new instance. KNN is chosen for its feature sensitivity, as slight variations in banknote features can differentiate a genuine note from a counter fit. Furthermore, KNN can effectively capture these subtle differences, especially with an appropriate distance function. KNN is also very interpretable, unlike the SVM classifier, and it can provide valuable insights and trends on how counterfeit bills are evolving.

Support Vector Machine (SVM)

The main objective of an SVM is to find an optimal hyperplane that separates different classes in the feature space. In the case of binary classification, this hyperplane acts as a decision boundary, dividing the data points into two distinct classes. The SVM aims to maximize the margin, the distance between the hyperplane, and the closest data points from each class. By maximizing the margin, SVMs achieve better generalization and improve their ability to classify unseen data correctly. SVM was chosen for its known abilities in classification problems and strong guarantees about its generalization abilities. SVM's use of kernel functions makes it very flexible depending on the shape of the data (linear/non-linear). It is also very good at dealing with imbalanced datasets, which could be a problem in banknote authentication as there might not be enough samples of counterfeit vs. samples of genuine bills.

Linear Kernel In SVMs, the linear kernel calculates the similarity between two feature vectors by taking their dot product. The result of this dot product represents the measure of similarity or correlation between the two vectors. The decision boundary generated by the linear kernel is a hyperplane in the feature space that aims to separate the classes by maximizing the margin between the closest data points from each class.

Gaussian Kernel (Radial Basis Function) The Gaussian kernel calculates the similarity between two data points based on their distance in the input space using the Gaussian or normal distribution. It assigns higher weights or similarity scores to data points closer to each other and decreases the weights as the distance increases. This creates a smooth and continuous transition of similarity values as we move away from a data point. The Gaussian kernel enables SVMs to capture non-linear relationships and model complex decision boundaries.

Gaussian Naive Bayes

Gaussian Naive Bayes is a classification algorithm that assumes the feature variables follow a Gaussian distribution. It estimates the mean and variance for each feature in each class during training and uses these estimates to calculate the likelihood of feature values for each class. Based on Bayes' theorem, it then calculates the posterior probability of each class given the observed feature values and assigns the class with the highest probability as the predicted class label. The features in the dataset seemed to be independent at the initial investigation, making Gaussian Naïve Bayes a possible candidate as it can perform very well with the dataset with independent variables. Naïve Bayes is also very efficient at eliminating irrelevant features due to its probabilistic nature, and some features in the dataset may not be equally informative.

These algorithms were chosen for their diverse characteristics and effectiveness in classification tasks. The performance of each algorithm was assessed by evaluating their accuracy, precision, recall, and confusion matrices using Cross-validation techniques to obtain reliable and robust performance estimates.

Variable Analysis

Variable Visualization

Variable plots are used to visualize one variable relationship. Scatter plots of each two-variable combination are used to visualize their relationships and potential discriminative power in distinguishing the bills. With the visualization, we may be able to notice the distribution and patterns of the variable combinations and pave the way to understanding the nature of the relationship between variables.

The combination of the variables can be expressed as such, consider all the attributes (except for class) as a Set u

u = Variance, Skewness, Curtosis, Entropy

Then all the two variable combinations of the attributes can be expressed as shown below

$s \in P(u) | cardinality(s) = 2$

Single Variable

The bi-modal shape of variance vs variance is illustrated in Figure 2. This suggests a large difference in the mean across separating different classes, and the correlation matrix of the variables shows a correlation coefficient of -0.72 between variance and class, indicating a strong linear inverse relationship. This could indicate that the variable "variance" is potentially informative and discriminatory in distinguishing between real and counterfeit banknotes. The variable entropy is shown to have two peaks representing different classes that overlap with each other, suggesting that there is significant ambiguity between classes based on their entropy values. This is also evident in the entropy correlation coefficient of -0.023 with the class variable, indicating a weak and negligible linear relationship. It suggests that machine learning algorithms are not able to establish clear decision boundaries based on entropy values.

Remaining variables: Skewness and curtosis's plot do not exhibit clear patterns or distinct separations between the classes, indicating potential challenges in using these individual variables for our model. The correlation coefficient between Skewness and class variable is -0.44, indicating a moderate negative linear relationship. While this suggests some degree of association, it is not as strong as one observed for variance. Lastly, the correlation coefficient of 0.16 between curtosis and class indicates a weak positive relationship. Characteristics of Skewness and Curtosis suggest limited discriminatory power when these variables are used individually to determine whether a banknote is genuine or counterfeit.

Based on these preliminary observations, it appeared that variable entropy could be potential noise among the variable considered for the banknote authentication task, as its scatter plots with other variable displays significant overlap and ambiguity. While variance and other variables are correlated with class, they alone do not have the discriminating power to consistently and accurately predict the class.

Two Variables

Two variable graphs are also illustrated in Figure 2. Out of all the combinations of variables, the scatter plot of curtosis vs entropy has the most significant overlap between classes. The overlapping of the data indicates that there is substantial ambiguity and similarity between classes based just on their curtosis and entropy values, which implies that there is no distinct separation that allows for easy classification. Machine learning algorithms would face significant challenges in accurately classifying banknotes based on these two variables. Aside from Curtosis vs Entropy, Skewness vs Entropy and Skewness vs Curtosis also have relatively high overlap between classes. The overlapping feature also has some implications for the effectiveness of the machine algorithms. A large overlap of different classes will require more complex decision boundaries to determine the class of the banknote accurately. Thus, it is reasonable to predict that algorithms such as Logistic Regression and Linear Discriminant Analysis, which assume linear relationships between variables, may struggle with an accurate prediction with only two features. Algorithms like KNN or SVM may be able to perform better with overlapping data by considering local patterns and neighbors.

Aside from Curtosis vs Entropy, Skewness vs Entropy and Skewness vs Curtosis also have relatively high overlap between classes. The overlapping feature also has some implications for the effectiveness of the machine algorithms. A large overlap of different classes will require more complex decision boundaries to determine the class of the banknote accurately. Thus, it is reasonable to predict that algorithms such as Logistic Regression and Linear Discriminant Analysis, which assume linear relationships between variables, may struggle with an accurate prediction with only two features. Algorithms like KNN or SVM may be able to perform better with overlapping data by considering local patterns and neighbors.

Three Variable Plot

As the combination of three variables can be expressed similarly to the two-variable expression but with a cardinality of 3, we extended the analysis by incorporating an additional feature, which resulted in three-feature scatter plots. By including a new variable, we aimed to mitigate the issue of significant overlap observed in the two feature scatter plot and improve the separability between the classes.

The three feature scatter plots exhibit a notable reduction in overlap among the different classes compared to the two feature plots, suggesting the additional variable added to any of the two variable combinations provided additional discriminatory power,

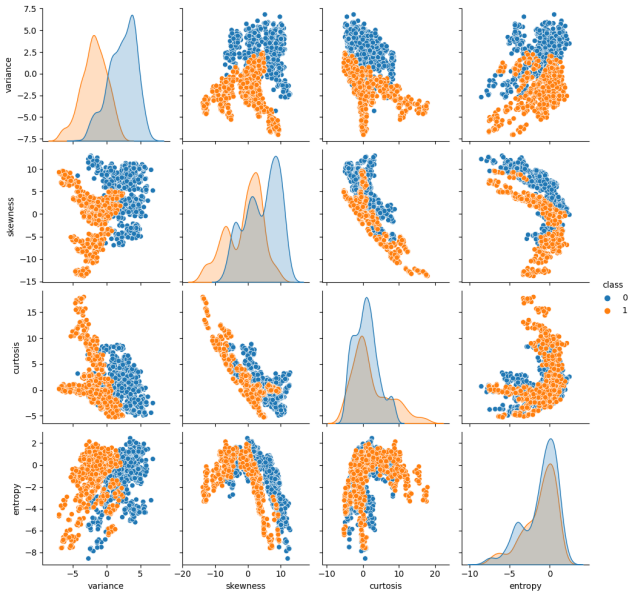


Fig. 2 One variable and two variable plots

leading to clearer separation between genuine and counterfeit banknotes. This improvement can be most clearly observed in Figure 3(d), where a single hyperplane or decision boundary can almost separate the points in different classes when variance, Skewness and curtosis were combined. The reduced overlap and improved separability in the three feature scatter plots indicate that machine learning algorithms, including those that assume linear relationships among the variables, such as Logistic Regression and Linear Discriminant Analysis, are expected to perform better when using three features for classification.

However, it is important to note that the above observation is based on the visual analysis of the scatter plots, and the investigation of variable interaction is still ongoing. Further quantitative analysis is necessary to confirm the extent of the visual analysis and its impact on classification accuracy.

Variable and Algorithm Selection

Wrapper Method Analysis and Feature Selection

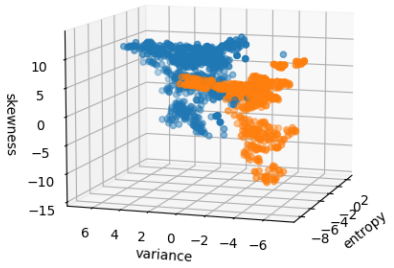
Having identified variable and variable sets with the most discriminative power for banknote authentication tasks and potential sources of noise within the dataset, we performed the wrapper method analysis, which provided valuable quantitative insights into the selection of variables and their impact on the performance of the Logistic Regression for banknote authentication.

Two Variable Wrapper Method Analysis Based on the result obtained from the Linear Regression Model with two different variable combinations, we observed that the feature combination involving variance has consistently demonstrated higher performance than other variable combinations (Table 2). The combination of variance and Skewness demonstrated the highest performance out of any other pairs, with an accuracy of 0.88 and strong precision, recall, and F1-score values, indicating significant discriminatory power and contributing significantly to the accurate classification of banknotes.

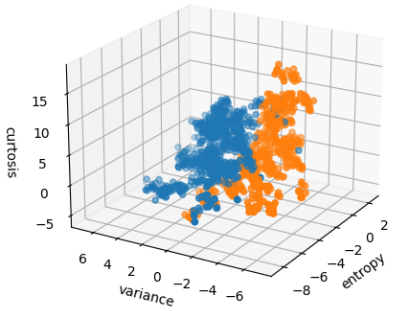
Another crucial observation is the notably lower recall value with the c-e feature pair, which suggests that Logistic Regression struggles significantly to identify positive instances for this featured pair. The confusion matrix (Figure 4) below reveals a higher number of false negatives with c-e feature pair, indicating that many positive instances are being misclassified as negatives. The combinations involving entropy showed lower performance measures, indicating their limited contribution to the classification task. This finding reaffirms our earlier theory that the variable of entropy may introduce noise or ambiguity, as indicated by its scatter plot and weak correlation with the class variable.

Three Variable Wrapper Method Analysis In addition to two variable combinations, we also performed a wrapper method analysis combining three variables using Logistic Regression. Among these three-variable combinations, the combination of Skewness, curtosis, and variance yielded the highest accuracy of 0.99, indicating this combination has excellent discriminatory power and performs well in distinguishing between genuine and counterfeit banknotes(Table 3). This reasserted our previous observation, where we noticed that the 3D scatter plot of Skewness, curtosis, and variance classes could be almost separated using a plane or a decision boundary; the results of the three variable wrapper method aligned with our previous findings, confirming the importance of Skewness, curtosis, and variance as informative features for banknote authentication and the separation in 3D scatter plot but not in 2D plots suggests that these variables collectively contribute to the accurate classification of banknotes.

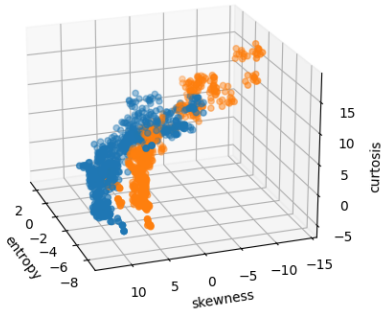
Although the highest accuracy combination does not contain variable entropy, it is worth noting that the selection of the three variables does not imply that the variable entropy is irrelevant, as three variable combinations that contain the variable entropy had the best accuracy of 0.91, which is still significant. Comparing those of two variable combinations, the confusion matrix (Figure 5) now shows much lower number of false negatives, even those of with entropy involved. Furthermore, our wrapper method analysis is specific to the logistic regression model, and other machine learning algorithms may yield different results. Thus, entropy still needs to be considered during our model selection process.



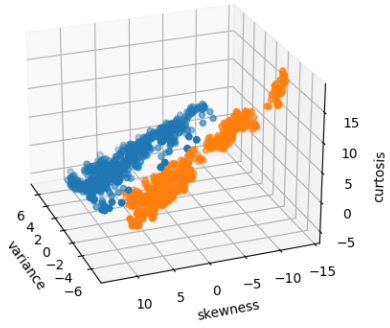
3a



3b



3c



3d

Fig. 3 Three variable plots

Table 2 Logistic Regression Two Variable Performance

Metric	v-s	v-c	v-e	s-c	s-e	c-e
accuracy	0.88	0.87	0.88	0.77	0.71	0.59
precision	0.88	0.84	0.87	0.73	0.71	0.59
recall	0.86	0.87	0.84	0.77	0.61	0.30
f1	0.87	0.86	0.86	0.75	0.66	0.40

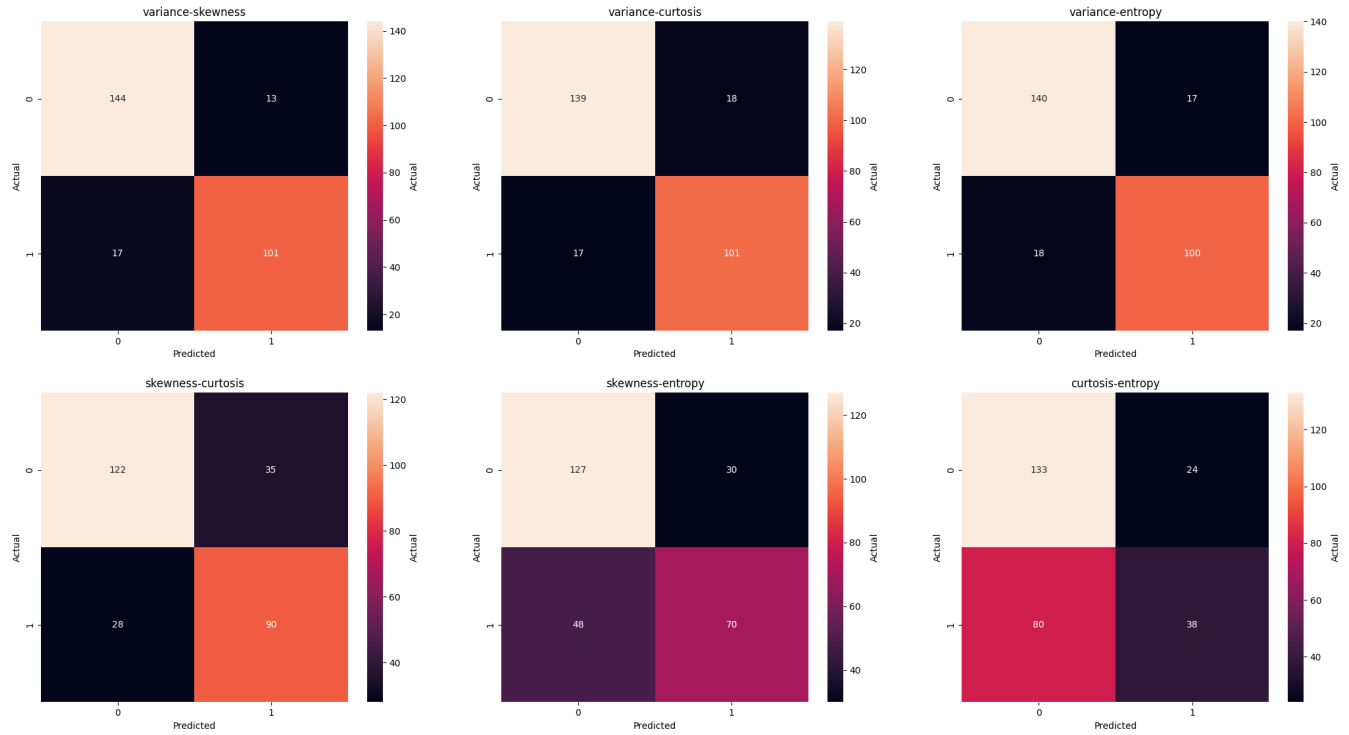


Fig. 4 Confusion Matrix Two variables

Table 3 Logistic Regression Three Variable Performance

Metric	v-s-c	v-s-e	v-c-e	s-c-e
accuracy	0.99	0.88	0.91	0.81
precision	0.98	0.87	0.89	0.77
recall	0.99	0.86	0.90	0.82
f1	0.99	0.86	0.90	0.79

Extended Variable Analysis

In section 4, we performed a visual analysis of the variables and noted the potential effect of the significant overlap of classes. We commented on the fact that algorithms that can operate complex decision boundaries may be more effective when only two variables are considered. Continuing the analysis, we further explored the performance of the KNN algorithm on the two-variable subsets. We ran the KNN algorithm with k = 5 with Euclidean distance for measuring the proximity between data points. The choice of k = 5 was made as a starting point to consider a reasonable number of neighbors without making the algorithm too sensitive to noise (k is too small) or too generalized (k is too large). Euclidean distance was chosen because it is a common and intuitive metric that measures the shortest line distance between two points in the feature space.

By comparing Table 4 with Table 2, it is clear that KNN outperformed Logistic Regression on two variable feature sets. For variance-skewness, KNN achieved an accuracy of 0.93 compared to 0.88 for Logistic Regression. The higher recall and f1-score for KNN also indicate that it is better at capturing true positives. And similar trends can be observed for the pair v-e, s-c, s-e, and v-c as well, suggesting a better overall classification performance.

For pair c-e, both algorithms performed relatively poorly compared to the other feature pairs. However, KNN still outperforms Logistic Regression with an accuracy of 0.71 compared to Logistic Regression's 0.59, and the ratio of false positive instances is significantly decreased as shown in KNN's confusion matrix in Figure 6.

The substantially higher performance of KNN and s-c feature pairs is especially noteworthy. As discussed earlier, this increase

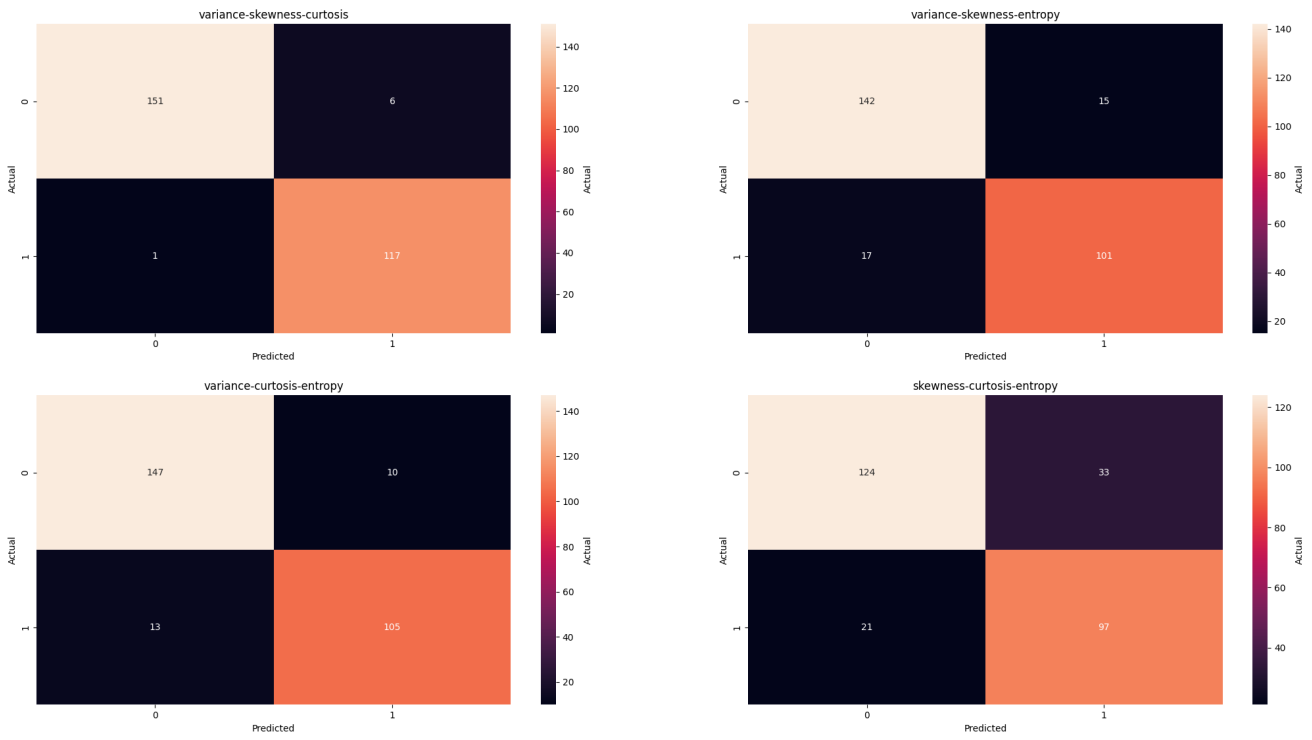


Fig. 5 Confusion Matrix: 3 Variables – KNN

Table 4 KNN Two Variable Performance

Metric	v-s	v-c	v-e	s-c	s-e	c-e
accuracy	0.93	0.89	0.90	0.91	0.88	0.71
precision	0.92	0.86	0.89	0.89	0.87	0.69
recall	0.93	0.91	0.90	0.89	0.88	0.65
f1	0.93	0.88	0.89	0.89	0.87	0.67

in performance could be attributed to KNN's ability to capture non-linear relationships. Logistic Regression, being a linear classifier, will likely struggle to effectively model the relationship between Skewness and Curtosis if they are non-linearly distributed or have complex interactions

Algorithm Comparison

After exploring the performance of the Logistic Regression and KNN algorithms on the two-variable and three-variable feature sets, we have identified the best-performing combination of variables. In this section, we investigated the performance of different Machine Learning algorithms on both the optimal feature sets and complete feature sets to ascertain which machine learning algorithm is best suited for banknote authentication. The algorithms considered for analysis include Logistic Regression, Linear Discrimination Analysis (LDA), K-Nearest Neighbors (KNN), Support Vector Machine (Linear Kernel), Support Vector Machine (Gaussian Kernel), and Gaussian Naive Bayes. We will consider the following feature sub-sets, Variance, Skewness, Curtosis, Variance, Curtosis, Entropy, as they had high performance during our wrapper method analysis and Variance, Skewness, Curtosis, Entropy which represents the complete set of features available in the dataset. It appears that there is no significant performance variation between the complete feature set and VCS feature set (Figure 7). The VCE feature set, however, showed a little bit lower accuracy across most algorithms tested. On the other hand, among the chosen algorithms, SVM with Gaussian Kernel (SVM-RBF) and KNN performed the best, and Naive Bayes was consistently lower than the other algorithms across all feature sets.

Table 5 Algorithm Comparison - VCSE Feature Set

Metric	LR	LDA	KNN	SVM-RBF	SVM-LINEAR	NB
accuracy	0.98	0.98	1.00	1.00	0.99	0.84
precision	0.96	0.95	1.00	1.00	0.97	0.84
recall	1.00	1.00	1.00	1.00	1.00	0.79
f1	0.98	0.97	1.00	1.00	0.98	0.81

Table 6 Algorithm Comparison - VCS Feature Set

Metric	LR	LDA	KNN	SVM-RBF	SVM-LINEAR	NB
accuracy	0.98	0.98	1.00	1.00	0.99	0.84
precision	0.96	0.95	1.00	0.99	0.97	0.84
recall	1.00	1.00	1.00	1.00	1.00	0.79
f1	0.98	0.97	1.00	1.00	0.98	0.81

Table 7 Algorithm Comparison - VCE Feature Set

Metric	LR	LDA	KNN	SVM-RBF	SVM-LINEAR	NB
accuracy	0.91	0.91	0.98	0.98	0.91	0.83
precision	0.89	0.89	0.97	0.96	0.89	0.83
recall	0.90	0.91	0.99	1.00	0.90	0.79
f1	0.90	0.90	0.98	0.98	0.90	0.81

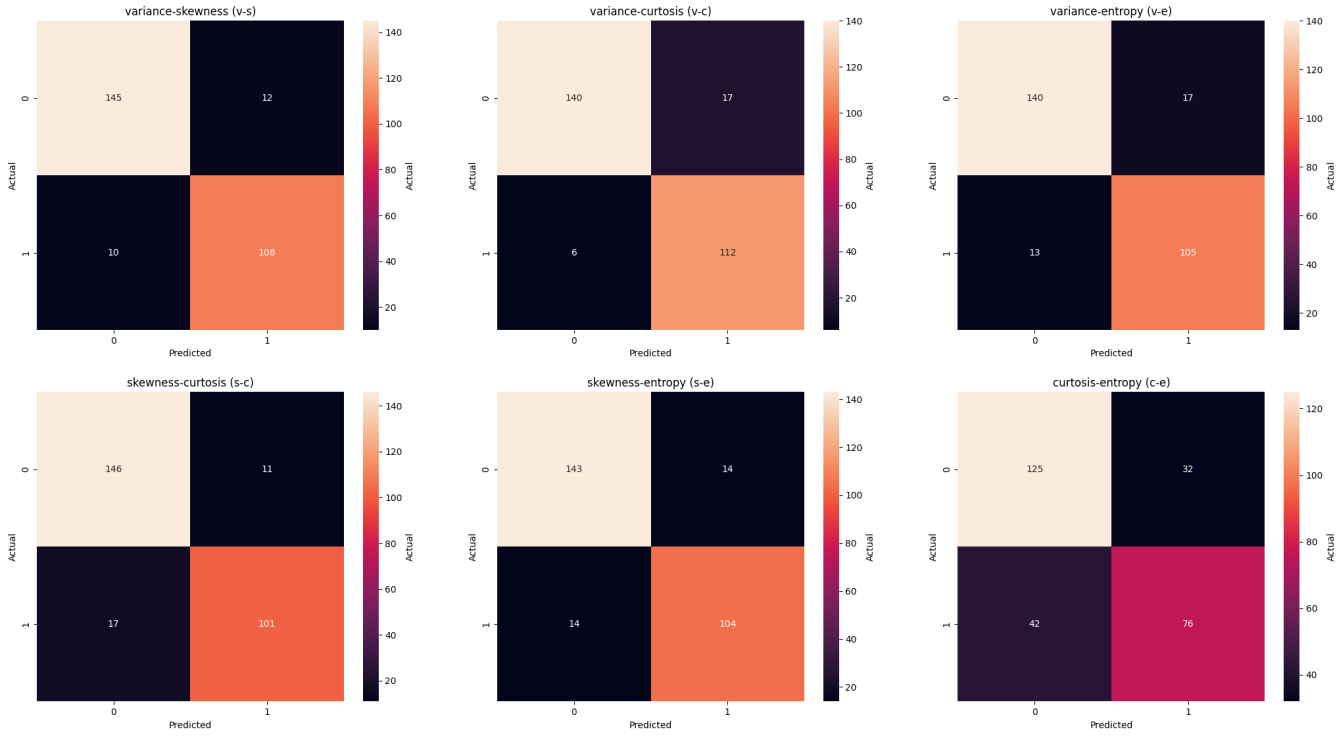


Fig. 6 Confusion Matrix: Two Variables – KNN

Table 5 presents the results using the complete feature set. The algorithm KNN and SVM with Gaussian Kernel (SVM-RBF) displayed the highest accuracy and F1 scores, closely followed by Logistic Regression and Linear Discriminant Analysis.

Table 6 reveals the results using the feature set VCS. Like the full feature set, KNN and SVM-RBF exhibit superior performance, and Logistic Regression and LDA have slightly lower accuracy.

Table 7 shows the results when using the feature set VCE. VCE's accuracy and F1 scores slightly dropped across most algorithms compared to the two previous feature sets.

The above observations suggest that feature subsets VCS can capture most of the information needed for accurately classifying banknotes. Entropy does not contribute significantly to the algorithm's performance compared to VCS and VCSE. However, we see a 0.01 performance increase with SVM-RBF with the Entropy variable, suggesting that there might be some marginal gain in the information captured by the Entropy feature in higher dimensions. On the other hand, when using the VCE feature set, there is a noticeable drop in accuracy across most algorithms, indicating that Skewness might be an essential feature for classification.

The latter consistently achieved better performance when comparing SVM with a linear kernel with SVM with an RBF kernel. This suggests that the relationship between features and class may not be linear.

The performance of Naive Bayes has been consistently lower than the other algorithms across all feature sets, indicating that it may not be suitable for banknote authentication. The low performance of Naive Bayes could be due to the dataset violating Naïve Bayes's assumptions.

Discussion

The comparative analysis in section 5 focuses on identifying the algorithm that exhibits the highest accuracy and overall performance in distinguishing between genuine and counterfeit bank notes. Additionally, we analyzed the impact of different variable subsets on the performance of each algorithm to gain insights into the interplay between variable selection and algorithm choice.

The results suggest that KNN and SVM with Gaussian kernel (SVM-RBF) are the most suitable for banknote authentication, as they consistently demonstrate high accuracy across different feature sets. KNN performed equally good regardless of the complete feature set or VCS; SVM-RBF achieved slightly better performance with the complete feature set than with VCS. The higher performance exhibited by KNN and SVM-RBF while

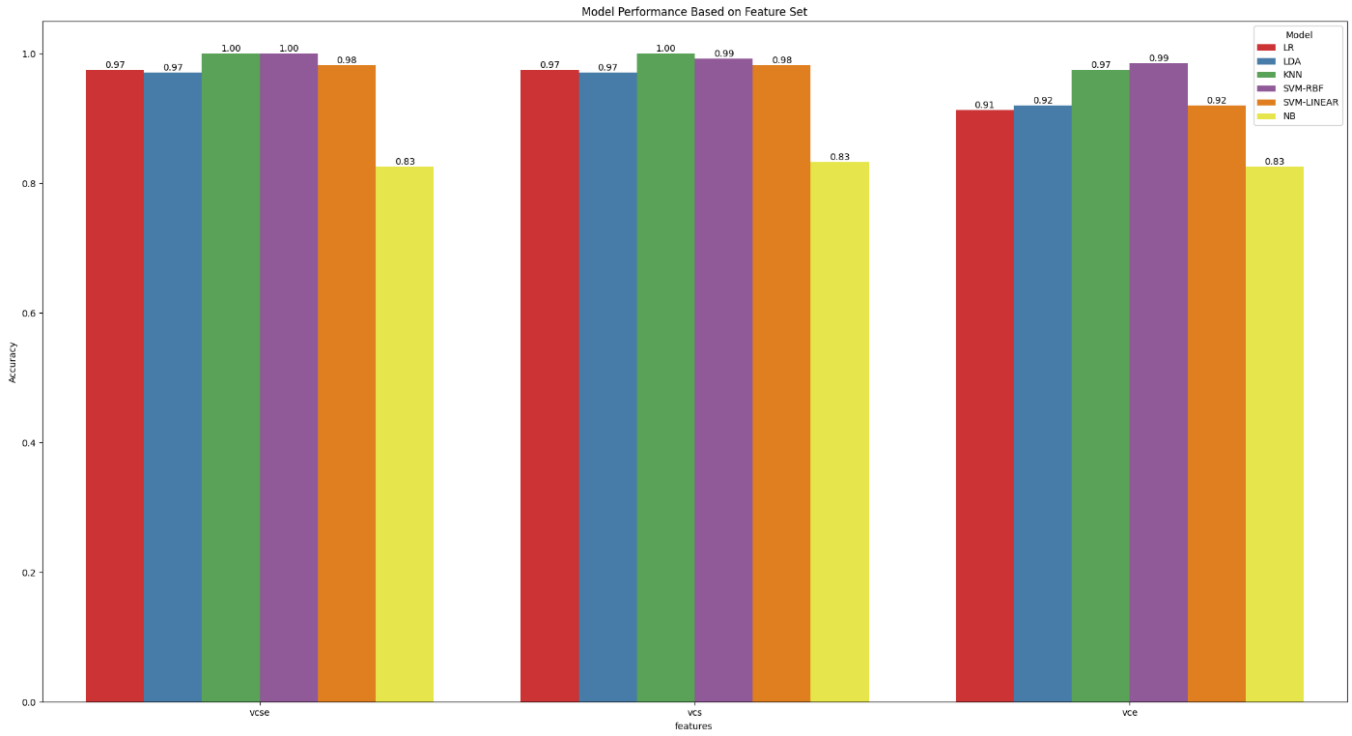


Fig. 7 Algorithm Comparison - Model Accuracy

slightly lower accuracy with Linear Regression and Linear Discriminate Analysis suggests that the data might have non-linear boundaries. SVM-RBF is adept at handling non-linear relationships in the data as it projects the input data onto a higher dimensional space, making it easier to find the hyperplane that separates the classes. The ability for it to handle non-linear data is likely why SVM-RBF outperforms linear classifiers and SVM-Linear, as the RBF kernel's flexibility allows it to capture complex non-linear relationships in the data. KNN's performance could be attributed to its instance-based nature, as it does not make strong assumptions of the underlying data distribution. Furthermore, given that the features can be particularly complex and dependent, KNN's ability to classify based on the proximity to known instances can be particularly effective.

Although both KNN and SVM-RBF achieved high accuracy, they have different performance characteristics that may impact their applications in this scenario. KNN relies on the entire dataset for prediction, meaning that for every banknote to be authenticated, the algorithm must compute the distance to every other note in the dataset, which can be computationally intensive, especially if the dataset is large. For large, real-time authentication systems where speed is crucial, KNN may pose scaling challenges. On the other hand, SVM has a computationally intensive training process, but once trained, the algorithm can use stored support vectors rather than the entire dataset to generate predictions quickly. These performance characteristics can influence the suitability of specific scenarios where authentication systems are applied.

It is also crucial to consider the tradeoff associated with incorporating additional features, as they can increase the model's dimensional complexity. Our prior analysis revealed a marginal difference in SVM-RBF's performance between VCSE and VCS feature sets. However, even a slight 0.01 difference in performance can have practical applications. In critical applications such as banknote authentication with high quantities of banknotes processed, even a minuscule percentage improvement can result in a significant uptick in the number of accurately classified banknotes. This means that SVM-RBF should be used with the complete feature set in order to achieve the best performance.

Using the same UCI dataset, several researches have explored its potential in banknote authentication. P V Kumar et al 7 have tested three supervised machine learning algorithms , Support Vector Machine(SVM), Decision Tree(DT) and K- Nearest Neighbor (KNN) under three different train test ratios 80:20, 70:30, and 60:40 and measured their performance based on various quantitative analysis parameters like Precision, Accuracy, Recall, MCC, F1-Score and others. They found KNN was the

best, giving around 100 % accuracy for a particular train test ratio (80:20), followed by Decision Tree (around 99% accuracy) and SVM(around 98% accuracy) , which are consistent with our test results. The authors did not elaborate if the train test ratios have any impact on the accuracy.P B V Rajarao et al 8 employed seven algorithms,KNN, Decision Tree, SVM, Random Forest, Logistic Regression, Nave Bayes, and LightGBM (light gradient-boosting machine) in their research. The order of accuracy of these algorithms were, LightGBM (100%), Decision Tree and Nave Bayes (both 97.5%), KNN and Random Forest (both 95%), SVM and Logistic Regression (both 40%). Contrary to our results, their Nave Bayes outperformed SVM. While many studies have delved into machine learning for banknote authentication there remains a gap in understanding the interplay between different algorithms and combinations of feature subsets. And few studies have conducted a comprehensive comparative analysis to determine the optimal mix for the highest accuracy.

Conclusion

Based on our comprehensive comparative analysis with various variable selection techniques and machine learning benchmarks, we have identified KNN and SVM-RBF as the most accurate algorithms. While Linear Regression and linear discriminate analysis did demonstrate solid performance, they were overshadowed by the capabilities of KNN and SVM-RBF. This underscores the significance of leveraging advanced machine-learning algorithms when the data needs intricate pattern recognition.

The subpar performance of Naïve Bayes highlights the importance of algorithm selection with the dataset's characteristics. In scenarios like banknote authentication, where the datasets manifest complex interactions between features, algorithms that rely on strong assumptions about data distribution and independence may suffer.

This research's border implication extends beyond identifying the best-performing algorithms. In a real-world context, these algorithms and datasets can be easily deployed in various mobile applications. In many regions, especially developing countries, mobile phone penetration has outpaced traditional banking and infrastructure development, while traditional authentication machines are expensive and challenging to distribute. In those communities, the informal economy is the primary source of livelihood, with transactions being predominantly cash-based, meaning that the risk of counterfeiting is high. Our research can be the foundation for developing cost-effective, widely accessible, and user-friendly mobile-based authentication tools tailored for those communities; as we can conclude in this case, SVM-RBF, although with high initial training cost, can excel in mobile applications due to its high accuracy and low prediction cost. In addition, the mobile application can be a valuable source of data as banknote counterfeiting is an evolving issue, and the ability of the algorithm to continuously evolve and learn from new patterns is needed to combat counterfeiting effectively. As such, online learning algorithms should be considered to ensure that authentication systems remain effective as new counterfeiting technology emerges.

List of abbreviations

Abbreviation	**Full name**
CNN	Convolutional Neural Networks
SVM	Support Vector Machine
SVM-RBF	Support Vector Machine- Radial Basis Function
KNN	K- nearest neighbors
LDA	Linear Discrimination Analysis (LDA)
LR	Logistic Regression
NB	Naïve Bayes
LightGBM	light gradient-boosting machine

References

1 J. W. Lee, H. G. Hong, K. W. Kim and K. R. Park, *Sensors*, 2017, **17**, 313.

2 S. Sabat and S. Muhamad, 2022.

3 D. P. V. Kumar, V. Akhila, B. Sushmitha and K. Anusha, *International Journal for Research in Applied Science and Engineering Technology*, 2022, **10**, 2328–2335.

4 S. Shinde, L. Wadhwa, D. G. Bhalke, N. Sherje, S. Naik, R. Kudale and K. Mohnani, *Multidisciplinary Science Journal*, 2024, **6**, 2024018–2024018.

5 F. Logic.

6 V. Lohweg, *banknote authentication*, 2012, `https://archive.ics.uci.edu/dataset/267`.

7 P. B. V. Rajarao, B. S. N. Murthy, K. Lavanya, K. C. Lakshmi, L. S. S. N. Praveen and G. S. Lokesh, *JOURNAL OF ALGEBRAIC STATISTICS*, 2022, **13**, 3680–3688.

8 M. Dirik, *Journal of Fuzzy Extension and Applications*, 2022, **3**, 302–312.

9 M.Rooks, *The Variance Of Pixel Values In Image Processing*, 2022, `https://www.icsid.org/uncategorized/what-is-variance-image-processing/`.

10 R. Shanmugam and R. Chattamvelli, *R. Chattamvelli*, 2016, pp. 89–110.

11 D.-Y. Tsai, Y. Lee and E. Matsuyama, *Journal of Digital Imaging*, 2008, **21**, 338–347.

12 J. Mccaffrey, *RPubs - Banknote Authentication*, `https://rpubs.com/haydenta/banknote-authentication`.

13 I. H. Sarker, *SN computer science*, 2021, **2**, 160.