|  |
| --- |
| A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection  M. JAVID AHAMED B.Tech  Dept of Artificial Intelligence and Data Science  Mepco Schlenk Engineering College  Sivakasi, TamilNadu, India  e-mail: javidfaaz@gmail.com  B. DHARUN KUMAR B.Tech  Dept of Artificial Intelligence and Data Science  Mepco Schlenk Engineering College  Sivakasi, TamilNadu, India  e-mail: dharunkumarbaskar@gmail.com    R. SAILESH B.Tech  Dept of Artificial Intelligence and Data Science  Mepco Schlenk Engineering College  Sivakasi, TamilNadu, India  e-mail: @gmail.com  Abstract— The use of credit cards for normal and online purchases has expanded dramatically in recent years due to developments in electronic commerce and communication technology. Nonetheless, the number of fraudulent credit card transactions has been steadily increasing, resulting in enormous annual losses for financial institutions. The evolution reducing these costs requires the adoption of efficient fraud detection algorithms, but this is difficult because most databases on credit cards are incredibly unbalanced. Moreover, obtaining credit with traditional machine learning techniques because of the way they are designed, which includes a static mapping of the input vector, card fraud detection is inefficient. As a result, they are unable to adjust to the changing purchasing habits of credit card users. This research suggests a hybrid data resampling technique combined with a neural network ensemble classifier as an effective way to identify credit card fraud. Adaptive boosting (AdaBoost) uses a long short-term memory (LSTM) neural network as the base learner to generate the ensemble classifier. The synthetic minority oversampling technique is used to achieve the hybrid resampling in the meantime modified closest neighbor technique (SMOTE-ENN). It is shown that the suggested strategy is effective employing real-world credit card transaction datasets that are accessible to the public. The following algorithms are used to benchmark the performance of the suggested method: LSTM, decision tree, multilayer perceptron (MLP), support vector machine (SVM), and classic AdaBoost. With a sensitivity and specificity of 0.996 and 0.998, respectively, the suggested LSTM ensemble beat the other algorithms, according to the experimental results, which also demonstrate that the classifiers performed better when trained on the resampled data.  Keywords- AdaBoost, credit card, data resampling, fraud detection, LSTM,SMOTE-ENN, machine learning. |

# Introduction *(Heading 1)*

The surge in e-commerce over the last ten years has led to a notable rise in the use of credit cards. Fraudulent transactions have been steadily rising as a result of credit card usage. The financial industry has been badly damaged by fraudulent credit card transactions.

According to a recent analysis, credit card theft cost the country roughly 27.85 billion dollars in 2018, a 16.2% increase from the 23.97 billion dollars lost in 2017. By 2023, it's predicted to reach 35 billion dollars. Effective fraud prevention and monitoring can lower these losses. Meanwhile, a number of credit card fraud detection systems have been developed using machine learning (ML). However, because of the class imbalance in the datasets, credit card fraud detection is still difficult to learn. While there are other issues that have impeded the detection of credit card fraud, the most significant one is the class imbalance. A problem known as class imbalance arises when datasets with an uneven class distribution are used in various real-world machine learning applications. One class (the majority class), for instance, has higher sample counts than the other class (the minority class). Because the number of legitimate transactions greatly exceeds that of fraudulent transactions, most credit card transaction datasets are unbalanced. When trained with balanced data, the majority of conventional machine learning algorithms function well. Because the algorithms are not made to take the class distribution into account, but rather the error rate, the skewed class distribution causes conventional ML algorithms to perform biasedly in favor of the majority class. As a result, compared to samples from the majority class, more examples from the minority class are misclassified.

Three categories can be used to categorize the approaches used in the literature to classify imbalanced data: data-level, algorithm-level, and hybrid techniques. By undersampling the majority class or oversampling the minority class, or occasionally both, data-level approaches typically produce a balanced dataset. By altering the classifier to focus more on the examples from the minority class, algorithm-level approaches seek to address the issue of class imbalance. Cost-sensitive learning strategies and ensemble learning are two instances of algorithm-level approaches.

In the meantime, both data-level and algorithm-level techniques are combined in the hybrid methods.

Several studies have put forth various approaches to address the issue of unequal class distribution in credit card fraud detection. For instance, Padmaja et al. suggested a fraud detection technique that removes extreme outliers from the minority class samples by employing k-reverse nearest neighbor (KRNN). Second, the dataset underwent hybrid resampling, which involves undersampling the majority class and oversampling the minority class. Several classifiers, such as the naïve Bayes, ID3 decision tree, and k-nearest neighbor (KNN) classifiers, were trained using the resampled data.

The suggested strategy outperformed conventional resampling techniques in terms of performance.

A light gradient boosting machine (LightGBM) was the basis of a credit card fraud detection technique presented by Taha and Malebary. The optimization algorithm that was employed to adjust the LightGBM's hyperparameters was based on Bayesian analysis. The method produced results with a precision of 97.34% and an accuracy of 98.40%. Additionally, Randhawa et al. investigated the effectiveness of a few common machine learning algorithms as well as hybrid classifiers, such as majority voting-based ensemble classifiers. The outcomes of the experiment demonstrate that the majority voting method performs exceptionally well in identifying fraudulent transactions.

Unbalanced data continues to be a challenge despite the many studies that have been proposed to address it, particularly in credit card fraud detection. Recurrent neural networks (RNN), like long short-term memory (LSTM) and gated recurrent units (GRU), have demonstrated great promise in modeling sequential data since the development of deep learning. Because traditional machine learning algorithms do not adjust to the changing shopping habits of credit card customers, they misclassify data when applied to fraud detection systems, which is why they have not been successful in detecting credit card fraud.

This study uses the LSTM neural network to tackle this issue and provide a solid solution that models the time series in credit card transactions. The idea behind this study is that, in some cases, it can be more advantageous to look at the full transaction sequence as opposed to just individual transactions. This is because credit card data can be time-modeled using a method that can identify even slight variations in the purchasing behavior of valid customers.

The creation of a reliable technique for detecting credit card fraud using an ensemble of long short-term memory neural networks is the study's contribution. Throughout, we apply an efficient feature engineering technique by employing the SMOTE technique to resample the unbalanced data. The LSTM neural network serves as the base learner for the adaptive boosting (AdaBoost) algorithm in the suggested ensemble technique. Two things make this approach noteworthy: first, the LSTM is a reliable algorithm for modeling sequential data. Second, the AdaBoost method creates robust classifiers with fewer false-positive predictions and a lower propensity to overfit.

Therefore, combining the AdaBoost algorithm with the LSTM neural network may be a great way to detect credit card fraud.

The remainder of this paper is organized as follows: The credit card fraud detection dataset and the traditional AdaBoost and LSTM methods are covered in Section II. The suggested credit card fraud detection system, complete with feature engineering and an LSTM ensemble, is presented in Section III. While Section V concludes the paper and offers recommendations for future research, Section IV presents the findings and discussions.

# BACKGROUND

DATASET:

# The popular credit card fraud detection dataset is used in this study. The Université Libre de Bruxelles (ULB) Machine Learning Group on big data mining and fraud detection prepared the dataset. The dataset includes credit card transactions made by European credit card clients in September 2013 that were completed in less than two days. Out of 284 807 transactions, only 492 are fraudulent, indicating an imbalance in the dataset. Due to the dataset's transformation, all of the attributes—aside from "Time" and "Amount"—are now numerical, and for privacy concerns, they are coded as V1, V2,..., V28. The transaction cost is represented by the "Amount" attribute, and the time interval in seconds between a transaction and the dataset's first transaction is represented by the "Time" attribute.

Lastly, the attribute ‘‘Class’’ is the dependent variable, and it has a value of 1 for fraudulent transactions and 0 for legitimate transactions.

ADAPTIVE BOOSTING:

Using an ensemble approach, the AdaBoost algorithm votes on the weighted predictions made by the weak learners in order to create strong classifiers. It has demonstrated exceptional performance in a number of applications, such as intrusion detection systems and credit card fraud detection.

In machine learning applications, overfitting frequently occurs, which results in subpar classification performance. Nonetheless, there is a decreased chance of overfitting and a higher likelihood of high false-positive predictions in classifiers trained with the AdaBoost methodology. An algorithm is chosen in the AdaBoost implementation to use the initial input data to train the base classifier.

Second, the sample weights are changed, giving the incorrectly classified samples more weight. Additionally, the next base learner, which aims to rectify the misclassifications from the earlier models, is trained using the adjusted instances. Until the desired number of models are constructed or all of the data's misclassified samples are removed, the iteration keeps going.

LONG SHORT-TERM MEMORY NEURAL NETWORK:

Extended Short-Term A unique variety of recurrent neural network (RNN) known as a memory neural network performs exceptionally well in learning long-term dependencies and stays clear of the gradient disappearance issue [24]. A memory cell CT is used in LSTMs to retain previous information, and three different kinds of gates regulate how the historical data is utilized and handled. The forget gate (ft), input gate (it), and output gate (ot) are the three gates. The following formulas are used to update the LSTM layers:

it = σ(Vixt + Wiht−1 + bi) (1)

ft = σ(Vf xt + Wf ht−1 + bf ) (2)

c˜t = tanh(Vcxt + Wcht−1 + bc) (3)

ct = ft ⊗ ct−1 + it ⊗ ˜ct (4)

ot = σ(Voxt + Woht−1 + bo) (5)

ht = ot ⊗ tanh(ct) (6)

In the meantime, ∗ can stand for f, i, or o to indicate the memory cell's particular gate or c. Thus, h∗ denotes the hidden state, b∗ is the bias, and ht is the output vector at time instant t. V∗ and W∗ are the weight matrices. Additionally, the sigmoid and tanh activation functions are represented by σ and tanh . The Hadamard or element-wise product is represented by the operator ⊗.

The LSTM algorithm starts by identifying any unnecessary data that should be eliminated from the cell. Depending on the decisions made by the input, output, and forget gates, respectively, an LSTM cell functions as a memory to write, read, and delete information .

# PROPOSED CREDIT CARD FRAUD DETECTION METHOD

SMOTE:

Synthetic Minority Over-sampling Technique, or SMOTE for short, is an algorithm used in machine learning to solve the issue of class imbalance, especially in classification tasks. When one class in the dataset has substantially fewer instances than the other class or classes, it is said to be imbalanced. Due to their propensity to favor the majority class, machine learning models may perform poorly in these situations.

The SMOTE algorithm is explained in general terms as follows:

* Determine Instances of Minority Classes:

Find the examples in your dataset that belong to the minority class.

* Choose an Example of a Minority Class:

Select an instance at random from the minority class to serve as the basis for creating artificial samples.

* Get the Closest Neighbors:

Determine the selected instance's k-nearest neighbors within the minority class. K's value is a parameter that the user defines.

* Create Synthetic Samples: In the feature space, draw a synthetic sample for each neighbor, joining the selected instance and its neighbor along the line. The degree of imbalance you wish to address will dictate how many synthetic samples you need to create.
* Again:

Until the appropriate balance between the classes is reached, repeat steps 2-4.

# Ease of Use

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this template and download the file for A4 paper format called “CPS\_A4\_format”.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”.

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the mTimes New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select Format > Columns >Presets > One Column.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

#### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select Format > Columns > Presets > One Column.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Format > Columns and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a Table footnote. (Table footnote)

We suggest that you use a text box to insert a graphic (ideally 300 dpi), with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord pull-down menu, select Format > Borders and Shading > Select “None”.

1. Example of a ONE-COLUMN figure caption.

Please see last page of this document for AN EXAMPLE of a 2-COLUMN Figure.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

## Footnotes

Use footnotes sparingly (or not at all) and place them at the bottom of the column on the page on which they are referenced. Use Times 8-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

# Copyright Forms and Reprint Orders

You must submit the Copyright Form per Step 7 of the CPS author kit’s web page. THIS FORM MUST BE SUBMITTED IN ORDER TO PUBLISH YOUR PAPER.

Please see Step 9 for ordering reprints of your paper. Reprints may be ordered using the form provided as <reprint.doc> or <reprint.pdf>.

##### Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R.B.G.) thanks . . .” Instead, try   
“R.B.G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

##### References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

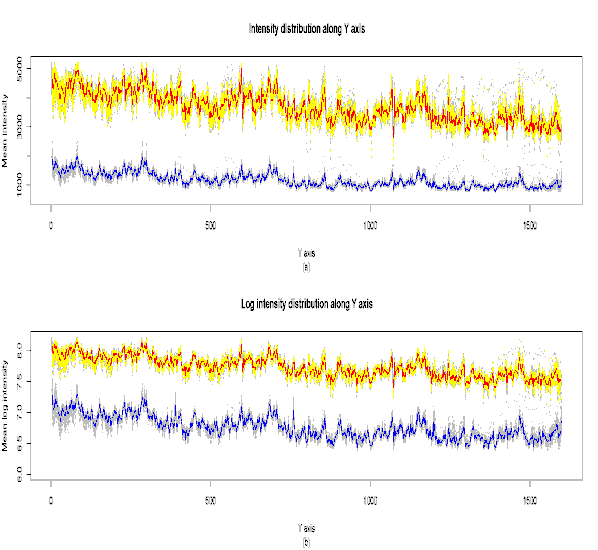
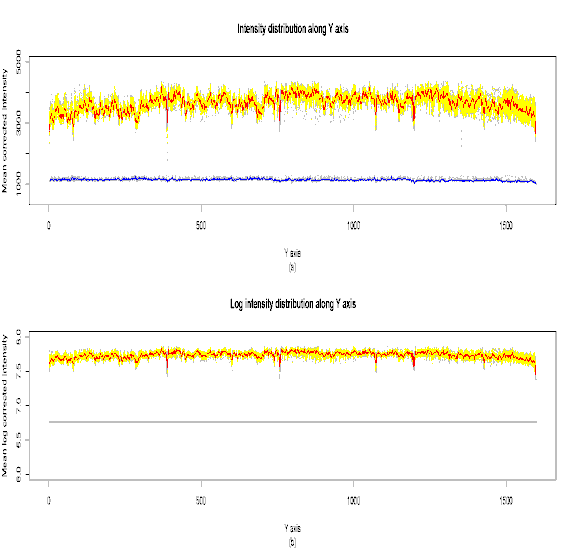
1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
8. Electronic Publication: Digital Object Identifiers (DOIs):

Article in a journal:

1. D. Kornack and P. Rakic, “Cell Proliferation without Neurogenesis in Adult Primate Neocortex,” Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

Article in a conference proceedings:

1. H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representatives,” Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

1. Example of a TWO-COLUMN figure caption: (a) this is the format for referencing parts of a figure.