Random Forests as a Tool for Effective Fraud Detection in Financial Systems

Sathvik Quadros[1], *Jacob Findley*[1]

[1]Dept of Computer Science, University of New Mexico, USA

sjquadros2004@unm.edu, jfindley@unm.edu

Abstract

Abstract - Financial systems must deal with hundreds of millions of transactions per day. While many of these are legitimate, a common problem financial systems must deal with is stopping adversaries attempting to commit fraud. Failing to detect fraud could lead to the loss of a significant amount of capital, and cause long lasting damage to the organizations affected. In this paper, we will discuss a machine learning approach to determine whether a specific transaction shows signs of being fraudulent. We implement and train a random forest consisting of many decision trees and evaluate their performance over several parameters such as attribute-selection criteria and confidence level. The experiments show that after all the training, the random forest tends to perform relatively well with an accuracy of 82.3%, given the inherent nature of fraudulent transactions being to disguise themselves as being legitimate.

Index Terms: random forest, decision tree, information gain, fraud detection, financial systems, artificial intelligence, machine learning

1. Introduction

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. It involves training a model on a large set of examples to extract generalized knowledge that can be extrapolated on. It requires a well specified task, a measure of performance, and plenty of training examples [1].

The classification of fraudulent and legitimate transactions is a well-specified task. Its measure of performance is the accuracy of its predictions. A large, but highly unbalanced set of instances is available, totaling at around 591k instances. 20% of these instances are reserved for the testing set, with the remaining 472k instances used for training. The large number of samples make machine learning a viable approach for solving the task of classification, and a random forest in particular is suitable as the data is represented as attribute-value pairs and the target function has discrete output values [1].

2. Design and Implementation

Authors are encouraged to describe how their work relates to prior work by themselves and by others, and to make clear statements about the novelty of their work.

2.1. Data Structures

All papers submitted to Odyssey 2026 must be original contributions that are not currently submitted to any other conference, workshop, or journal, nor will be submitted to any other conference, workshop, or journal during the review process of Odyssey 2026. Cross-checks with submissions from other conferences will be carried out to enforce this rule.

2.2. ID3 Algorithm

All (co-)authors must be responsible and accountable for the work and content of the paper, and they must consent to its submission. Generative AI tools cannot be a co-author of the paper. They can be used for editing and polishing manuscripts, but should not be used for producing a significant part of the manuscript.

Algorithm 1 id3(An algorithm with caption)

Require: $n \geq 0$
Ensure: $y = x^n$
  $y \leftarrow 1$
  $X \leftarrow x$
  $N \leftarrow n$
  while $N \neq 0$ **do**
    if N is even **then**
      $X \leftarrow X \times X$
      $N \leftarrow \frac{N}{2}$ ▷ This is a comment
    else if N is odd **then**
      $y \leftarrow y \times X$
      $N \leftarrow N - 1$
    end if
  end while

2.3. Attribute Selection

The theme of Odyssey 2026 is *Speech beyond words: Trustworthy Identity, Health, Emotion and more*. Odyssey 2026 continues to be fully committed to advancing speech science and technology while meeting new challenges. Please refer to the conference website for further detail.

2.4. Overfitting and Branch Pruning

The theme of Odyssey 2026 is *Speech beyond words: Trustworthy Identity, Health, Emotion and more*. Odyssey 2026 continues to be fully committed to advancing speech science and technology while meeting new challenges. Please refer to the conference website for further detail.

2.5. Classification

The theme of Odyssey 2026 is *Speech beyond words: Trustworthy Identity, Health, Emotion and more*. Odyssey 2026 continues to be fully committed to advancing speech science and

technology while meeting new challenges. Please refer to the conference website for further detail.

3. Experiments

For all the experiments we decided on a further 90/10 split of the instances selected for training, using 10% of them instances for validation.

The page layout should match with the following rules. A highly recommended way to meet these requirements is to use one of the templates provided and to check details against this example file. Do not modify the template layout! Do not reduce the line spacing!

If for some reason you cannot use any of the templates, please follow these rules as carefully as possible, or contact the organizers at <info@odyssey2026.org> for further instructions.

3.1. Basic layout features

- Proceedings will be printed in A4 format. The layout is designed so that the papers, when printed in US Letter format, will include all material but the margins will not be symmetric. PLEASE TRY TO MAKE YOUR SUBMISSION IN A4 FORMAT, if possible, although this is not an absolute requirement.
- Two columns are used except for the title part and possibly for large figures that may need a full page width.
- Left margin is 20 mm.
- Column width is 80 mm.
- Spacing between columns is 10 mm.
- Top margin is 25 mm (except for the first page which is 30 mm to the title top).
- Text height (without headers and footers) is maximum 235 mm.
- Page headers and footers must be left empty.
- No page numbers.
- Check indentations and spacing by comparing to the example PDF file.

3.2. Section headings

Section headings are centred in boldface with the first word capitalised and the rest of the heading in lower case. Sub-headings appear like major headings, except they start at the left margin in the column. Sub-sub-headings appear like sub-headings, except they are in italics and not boldface. See the examples in this file. No more than 3 levels of headings should be used.

4. Fonts

The font used for the main text is Times. The recommended font size is 9 points which is also the minimum allowed size. Other font types may be used if needed for special purposes. Remember, however, to embed all the fonts in your final PDF file!

LaTeX users: DO NOT USE THE Computer Modern FONT FOR TEXT (Times is specified in the style file). If possible, make the final document using POSTSCRIPT FONTS since, for example, equations with non-PS Computer Modern are very hard to read on screen.

4.1. Figures

Figures must be centred in the column or page (if the figure spans both columns). Figures which span 2 columns must be placed at the top or bottom of a page. Captions should follow each figure and have the format used in Fig. 1.

Figures should preferably be line drawings. If they contain gray levels or colors, they should be checked to print well on a high-quality non-color laser printer. If some figures contain bitmap images, please ensure that their resolution is high enough to preserve readability.

4.2. Tables

An example of a table is shown in Table 1. Somewhat different styles are allowed according to the type and purpose of the table. The caption text may be above or below the table. Tables must be legible when printed in monochrome on A4 paper.

Table 1: *This is an example of a table*

Ratio	Decibels
$1/10$	-20
$1/1$	0
$2/1$	≈ 6
$3.16/1$	10
$10/1$	20

4.3. Equations

Equations should be placed on separate lines and numbered. Examples of equations are given below. Particularly,

$$x(t) = s(f_\omega(t)) \tag{1}$$

where $f_\omega(t)$ is a special warping function

$$f_\omega(t) = \frac{1}{2\pi j} \oint_C \frac{\nu^{-1k} d\nu}{(1-\beta\nu^{-1})(\nu^{-1}-\beta)} \tag{2}$$

A residue theorem states that

$$\oint_C F(z)dz = 2\pi j \sum_k Res[F(z), p_k] \tag{3}$$

Applying (3) to (1), it is straightforward to see that

$$1 + 1 = \pi \tag{4}$$

4.4. Page numbering

Final page numbers will be added later to the document electronically. *Please do not include any headers or footers!*

4.5. Style

Manuscripts must be written in English. Either US or UK spelling is acceptable (but do not mix them).

4.5.1. References

It is ISCA policy that papers submitted should refer to peer-reviewed publications. References to non-peer-reviewed publications (including public repositories such as arXiv, Preprints, and HAL, software, and personal communications) should only

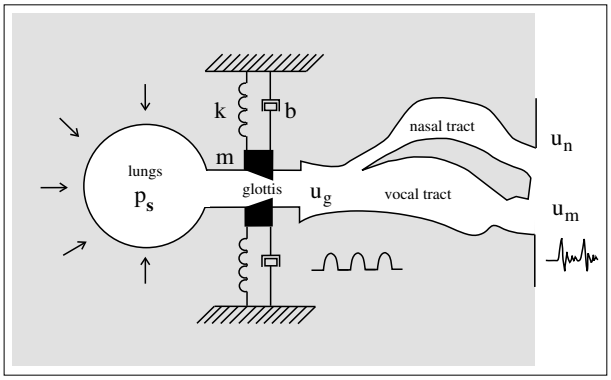

Figure 1: *Schematic diagram of speech production.*

be made if there is no peer-reviewed publication available, should be kept to a minimum, and should appear as footnotes in the text (i.e., not listed in the References).

References should be in standard IEEE format, numbered in order of appearance, for example [1] is cited before [2]. For longer works such as books, provide a single entry for the complete work in the References, then cite specific pages [3, pp. 417–422] or a chapter [3, Chapter 2]. Multiple references may be cited in a list [4, 5].

4.5.2. International System of Units (SI)

Use SI units, correctly formatted with a non-breaking space between the quantity and the unit. In LaTeX this is best achieved using the `siunitx` package (which is already included by the provided LaTeX class). This will produce 25 ms, 44.1 kHz and so on.

5. Submissions

Information on how and when to submit your paper is provided on the conference website.

5.1. Manuscript

Authors are required to submit a single PDF file of each manuscript. The PDF file should comply with the following requirements: (a) no password protection; (b) all fonts must be embedded; and (c) text searchable (do ctrl-F and try to find a common word such as "the"). The conference organisers may contact authors of non-complying files to obtain a replacement. Papers for which an acceptable replacement is not provided in a timely manner will be withdrawn.

5.1.1. Embed all fonts

It is *very important* that the PDF file embeds all fonts! PDF files created using LaTeX, including on `https://overleaf.com`, will generally embed all fonts from the body text. However, it is possible that included figures (especially those in PDF or PS format) may use additional fonts that are not embedded, depending how they were created.

On Windows, the bullzip printer can convert any PDF to have embedded and subsetted fonts. On Linux & MacOS, converting to and from Postscript will embed all fonts:

```
pdf2ps file.pdf
ps2pdf -dPDFSETTINGS=/prepress file.ps file.pdf
```

6. Discussion

Authors must proofread their PDF file prior to submission, to ensure it is correct. Do not rely on proofreading the LaTeX source or Word document. **Please proofread the PDF file before it is submitted.**

7. Acknowledgements

The Odyssey 2026 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for kindly providing the previous version of this template.

8. References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[4] J. Smith, F. Lastname2, and F. Lastname3, "A really good paper about Dynamic Time Warping," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 100–104.

[5] R. Jones, F. Lastname2, and F. Lastname3, "An excellent paper introducing the ABC toolkit," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 105–109.