Clearing Up Confusion: A Categorical Data Analysis Approach to Confusion Matrices

Kiran K. Bumra

220460358

PSYC 6136

Dr. Michael Friendly

May 7, 2026

Introduction

Confusion matrices are a fundamental tool for evaluating classification and diagnostic systems, widely used across fields like medicine, epidemiology, machine learning, and psychology. Foundationally, confusion matrices summarize the relationship between actual observed outcomes and predicted classifications, typically in the form of a cross-classified table (Lovell et al., 2021). For example, confusion matrices can describe the agreement between a diagnostic test and true disease status. Or, in a machine learning context, they can quantify the performance of predictive algorithms (Powers, 2011). Despite its versatility, confusion matrices are most often presented as performance evaluation tools, with emphasis placed on derived measures such as accuracy, sensitivity, and specificity, rather than as objects of statistical analysis in their own right.

From the perspective of Categorical Data Analysis (CDA), confusion matrices can be understood more fundamentally as square contingency tables representing the joint distribution of two categorical variables: the actually observed class and the predicted class. This perspective places confusion matrices within a well-established statistical framework, where methods for analyzing association, agreement, and dependence between categorical variables have been extensively developed (Agresti, 2013). However, this connection is rarely made explicit in applied contexts, where the focus tends to remain on predictive performance metrics without deeper consideration of the underlying structure of the data.

Viewing confusion matrices through a CDA lens offers several advantages. First, it allows standard tools such as odds ratios, chi-square tests, and measures of agreement (e.g., Cohen's kappa) to be applied directly to classification problems (Johnson & Johnson, 2014; Agresti, 2013). Second, confusion matrices can be used as a naturally extended framework to understand beyond the simple 2x2 case, including multiclass classification problems and issues arising from imbalanced data (Lovell et al., 2021). Finally, it connects modern applications in predictive modeling with classical statistical theory, offering a more interpretable approach to evaluating classification systems.

This paper aims to situate confusion matrices within the CDA framework by demonstrating their equivalence to square contingency tables and exploring their statistical properties and extensions. Specifically, the paper examines the 2x2 confusion matrix and its relationship to measures of association and diagnostic accuracy, reviews graphical approaches to

representing such tables, and extends the discussion to multiclass confusion matrices and the challenges posed by imbalance classifications (Lovell et al., 2021). Through these connections, the paper highlights how confusion matrices can be analyzed not only as performance summaries, but as structured categorical data flexible to the full range of methods in CDA.

The 2x2 Confusion Matrix as a Contingency Table

The simplest and most widely used form of a confusion matrix is the 2x2 table, which arises in binary classification problems where observations are categorized into one of two possible classes (e.g., positive versus negative). In its standard form, the confusion matrix cross-classifies the observed (actual) condition with the predicted classification, resulting in four possible outcomes: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). This structure is identical to a 2x2 contingency table, where the cell counts represent the joint frequencies of two categorical variables (Agresti, 2013). Figure 1 illustrates the basic structure of a 2x2 confusion matrix, showing how the four cell types map onto the joint distribution of observed and predicted classifications. Please note that all figures in this paper were generated in R to illustrate the categorical data structures and visualization methods discussed throughout the paper.

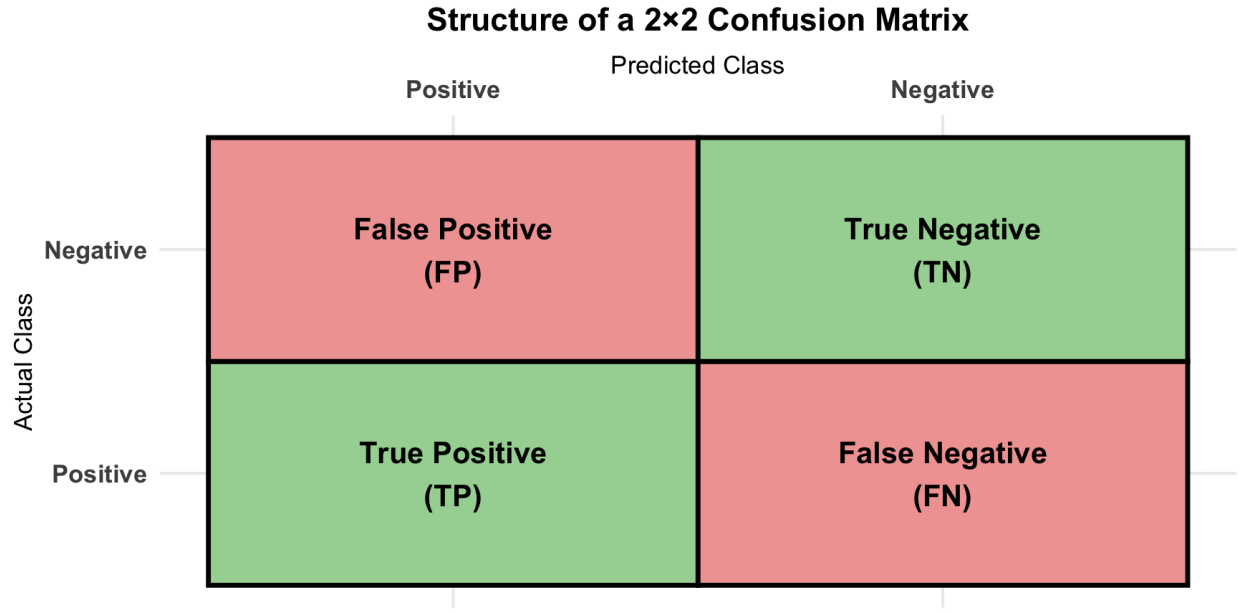

Figure 1. Structure of a 2x2 confusion matrix. Green cells represent correct classifications; red cells represent misclassification errors.

The confusion matrix can be viewed as the joint distribution of (Y, $\hat{Y}$) with Y representing the observed class and $\hat{Y}$ representing the predicted class with values in $\{0, 1\}$. The cell count (n_{ij}) represents the number of observations with true class (i) and predicted class (j). So, with this interpretation, the confusion matrix is a representation of the relationship between two categorical variables. This perspective aligns with the standard contingency table framework, in which inference focuses on assessing independence and association between categorical variables (Agresti, 2013).

Many of the performance measures commonly reported in diagnostic testing and classification studies can be expressed directly as functions of the cell counts in this table. Sensitivity, defined as the proportion of actual positives correctly identified (TP/(TP + FN)), and specificity, defined as the proportion of actual negatives correctly classified (TN/(TN + FP)), are fundamental measures in the evaluation of diagnostic tests (Pepe, 2003). These quantities correspond to conditional probabilities of correct classification given the true class, and are essential to the analysis of diagnostic procedures. Other measures such as precision (TP/(TP + FP)), and accuracy ((TP + TN)/(TP + FP + FN + TN)), are derived directly from the same joint frequency structure (Altman & Bland, 1994a; Altman & Bland, 1994b). Together, these measures illustrate that commonly used classification metrics are functions of the underlying contingency table.

From a CDA perspective, these performance measures can be complemented by classical measures of association. In particular, the odds ratio provides a natural summary of the strength of association between observed and predicted classifications. For a 2x2 table, the odds ratio is defined as (TP x TN)/(FP x FN), and quantifies the extent to which correct classifications are associated with the observed outcomes. Values greater than one indicate positive association between predicted and observed categories, while a value of one corresponds to statistical independence (Agresti, 2013). This interpretation reframes classification performance in terms of association between two categorical variables.

In addition, the chi-square test of independence can be applied to the confusion matrix to assess whether predicted and actual classifications are statistically related. Under the null hypothesis of independence, the predicted labels provide no information about the true labels, implying that the classification system performs no better than random chance. A significant

chi-square statistic therefore indicates that the classifier captures structure in the data beyond random variation, consistent with standard inference for contingency tables (Agresti, 2013).

Viewing the 2x2 confusion matrix as a square contingency table emphasizes its dual role as a performance summary and a statistical object. While commonly reported metrics such as sensitivity and specificity emphasize practical aspects of classification, measures such as the odds ratio and chi-square statistic provide insight into the underlying association between predicted and observed outcomes. This dual perspective forms the basis for extending confusion matrices to more complex settings, including multiclass classification and imbalanced data.

Measures of Agreement and Beyond Accuracy

While measures of sensitivity, specificity, and accuracy are commonly used to summarize classification performance, they do not fully capture the relationship between observed and predicted classifications. In particular, accuracy, defined as the proportion of correctly classified observations, can be misleading, especially in situations where class distributions are highly imbalanced (Altman & Bland, 1994a; Sokolova & Lapalme, 2009). For example, in a diagnostic setting where the prevalence of a condition is low, a classifier that predicts all observations as negative may achieve high accuracy despite having no ability to detect true positive cases (Pepe, 2003). This limitation highlights the need for measures that account not only for correct classification, but also for the agreement between observed and predicted outcomes beyond what would be expected by chance.

From the perspective of CDA, this issue can be addressed through measures of agreement that explicitly incorporate the role of marginal distributions. In contingency tables, the marginal totals strongly influence the level of agreement that can occur under independence, making it essential to distinguish between observed agreement and agreement expected by chance (Agresti, 2013). One of the most widely used statistics in this context is Cohen's kappa, which provides a chance-corrected measure of agreement for categorical data. Unlike accuracy, which treats all agreement as equally informative, kappa distinguishes between agreement that occurs due to the underlying distribution of categories and agreement that reflects a genuine relationship between variables (Cohen, 1960; Agresti, 2013). Cohen's kappa is defined as the proportion of observed agreement minus the proportion of agreement expected by chance, divided by the maximum possible agreement beyond chance (Cohen, 1960). In the context of a confusion matrix, the observed agreement corresponds to the portion of cases along the diagonal (i.e., (TP + TN)/N),

while the expected agreement is computed from the product of the marginal proportions of the observed and predicted classifications (Agresti, 2013). This adjustment is particularly important when the marginal distributions are uneven, as is often the case in real-world classification problems (Pepe, 2003). Figure 2 makes this divergence explicit by showing that as class imbalance increases, overall accuracy can rise because the majority class dominates the data, while Cohen's kappa provides a chance-corrected view of agreement that is less directly inflated by the marginal distribution. This illustrates that accuracy and chance-corrected agreement capture different aspects of classification performance. Figure 3 reinforces this point directly: two confusion matrices with identical accuracy of 90% produce substantially different kappa values simply due to differences in the marginal distribution of classes, demonstrating that accuracy alone is insufficient to evaluate classification performance.

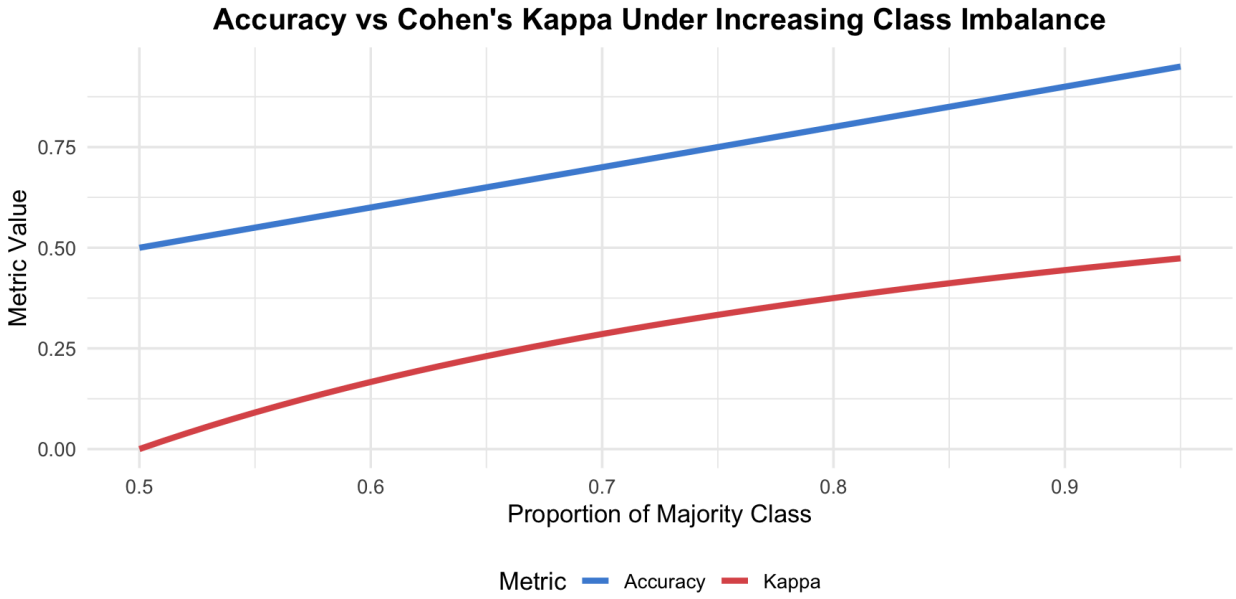

Figure 2. Accuracy and Cohen's kappa as a function of increasing class imbalance. As the proportion of the majority class increases, overall accuracy rises. Cohen's kappa, however, provides a chance-corrected measure of agreement and therefore shows that accuracy alone may overstate classification performance under imbalanced class distributions.

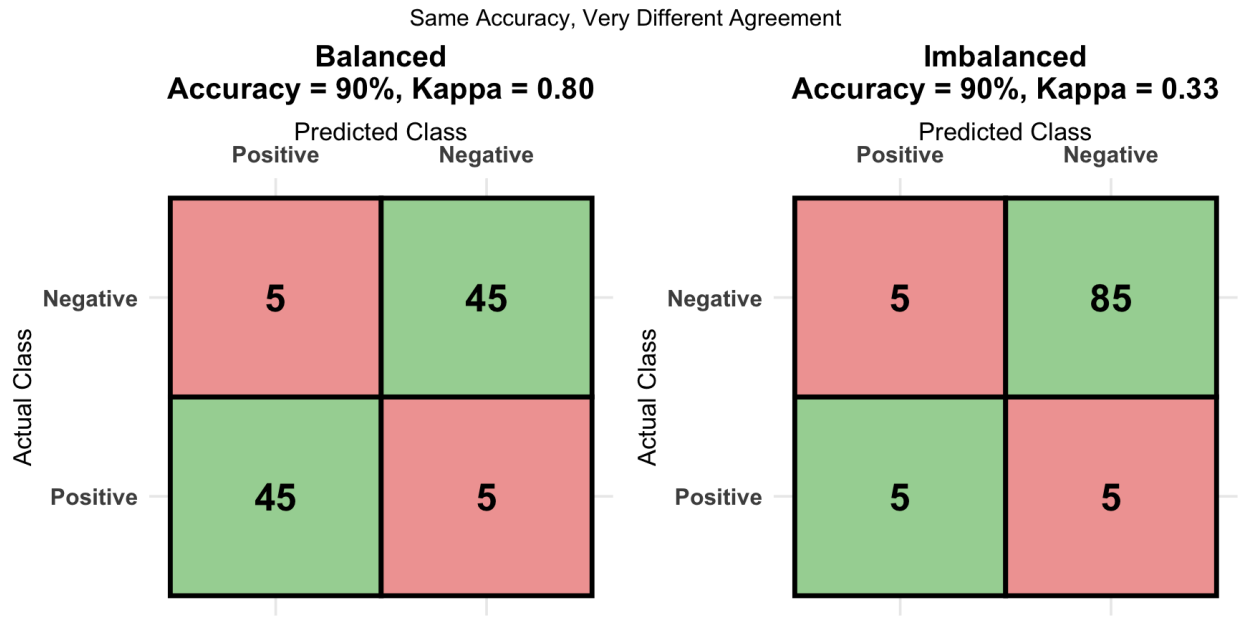

Figure 3. Side-by-side comparison of balanced and imbalanced 2x2 confusion matrices, each achieving 90% accuracy. Despite identical accuracy, Cohen's kappa reveals substantially different levels of agreement beyond chance, illustrating the limitations of accuracy as a sole performance metric.

The distinction between accuracy and chance-corrected agreement is critical for interpreting confusion matrices as contingency tables. In a 2x2 table with highly imbalanced margins, high accuracy may arise simply because one category dominated both the observed and predicted distributions, rather than because of meaningful predictive performance (Sokolova & Lapalme, 2009; Powers, 2011). In such cases, the expected agreement under independence is already large, and the incremental information provided by the classifier may be minimal. Cohen's kappa corrects for this by effectively comparing the observed joint distribution to what would be expected if the two variables were statistically independent (Cohen, 1960; Agresti, 2013).

This perspective highlights a deeper connection between confusion matrices and classical methods in categorical data analysis. Measures of agreement, such as kappa are closely related to models of independence and association in contingency tables, where the structure of the joint distribution is of primary interest (Agresti, 2013). While the chi-square statistic tests for the presence of association, kappa provides a normalized measure of agreement relative to chance, offering a complementary interpretation of the same underlying table (Cohen, 1960; Agresti,

2013). In this sense, confusion matrices can be analyzed not only through descriptive performance metrics, but also through the lens of statistical models that quantify agreement structure.

Extensions of kappa further reinforce this connection. For example, weighted kappa allows different types of disagreement to be penalized unequally, which is particularly useful when categories are ordered (Agresti, 2013). Although binary classification does not require weighting, this extension becomes important in multiclass settings, where some misclassifications may be more severe than others. More broadly, the framework of agreement analysis provides a principled way to evaluate classification systems while accounting for the structure of categorical data.

Overall, incorporating measures of agreement into the analysis of confusion matrices shifts the focus from raw predictive performance to the underlying relationship between observed and predicted classifications. While accuracy and related metrics provide useful summaries, they can obscure important features of the data, particularly in the presence of imbalance or skewed marginal distributions (Pepe, 2003; Powers, 2011). In contrast, agreement-based measures such as Cohen's kappa offer a more nuanced and statistically grounded interpretation, aligning confusion matrices more closely with the broader framework of CDA (Agresti, 2013).

Multiclass Confusion Matrices and Imbalanced Classifications

While the 2x2 confusion matrix provides a useful framework for binary classification, many real-world problems involve more than two categories. In these cases, the confusion matrix generalizes to a k x k table, where k represents the number of classes. Each cell (n_{ij}) represents the number of observations belonging to the true class (i) that are classified into category (j). As in the binary case, the diagonal elements correspond to correct classifications, while the off-diagonal elements represent misclassification errors. This structure can again be interpreted as a square contingency table describing the joint distribution of observed and predicted categorical variables, extending the framework of CDA to multiclass settings (Agresti, 2013).

Despite this generalization, the interpretation of multiclass confusion matrices is significantly more complex than in the binary case. One challenge is that many commonly used performance measures do not extend naturally or interpretably to higher dimensions. For example, overall accuracy remains easy to compute but becomes increasingly difficult to

interpret, particularly when the class distribution is uneven or when errors are not uniformly distributed across categories (Sokolova & Lapalme, 2009). In such cases, accuracy may obscure important patterns of misclassification, providing an incomplete or even misleading summary of model performance.

A central issue in multiclass classification is the presence of class imbalance, where some categories occur much more frequently than others. The problem has been widely recognized in both statistical and machine learning contexts, as imbalanced data can strongly influence both model estimation and evaluation (He & Garcia, 2009). In confusion matrices, imbalance manifests in highly unequal marginal totals, which in turn affect both observed agreement and expected agreement under independence. As a result, measures such as accuracy or even kappa may be dominated by the performance of the majority class, while performance on minority classes – often of greatest practical interest – may be poorly captured (Lovell et al., 2021).

To address these challenges, it is often useful to decompose a multiclass confusion matrix into a set of binary classification problems. One common approach is the one-versus-all framework, in which each class is treated as the positive class in turn, with all other classes combined into a single negative category. This decomposition allows binary performance measures such as sensitivity, specificity, and precision to be computed for each class separately, providing a more detailed view of classification performance (Sokolova & Lapalme, 2009). From a contingency table perspective, this corresponds to constructing a series of 2x2 tables derived from the original k x k table.

Building on this idea, Lovell and colleagues (2021) emphasize that multiclass confusion matrices can be systematically analyzed through pairwise comparisons between categories. Specifically, they show that the information contained in a multinomial confusion matrix can be understood in terms of a collection of binary subproblems, each corresponding to a pair of classes. This approach provides a more interpretable and granular analysis of classification performance, particularly in settings with severe class imbalance. Figure 4 extends this approach visually, using a mosaic display with residual shading to reveal which class pairs are most frequently confused and where classification performance deviates most from what would be expected under independence. By focusing on pairwise distinctions, it becomes possible to identify which categories are most frequently confused and to assess the relative difficulty of different classification boundaries.

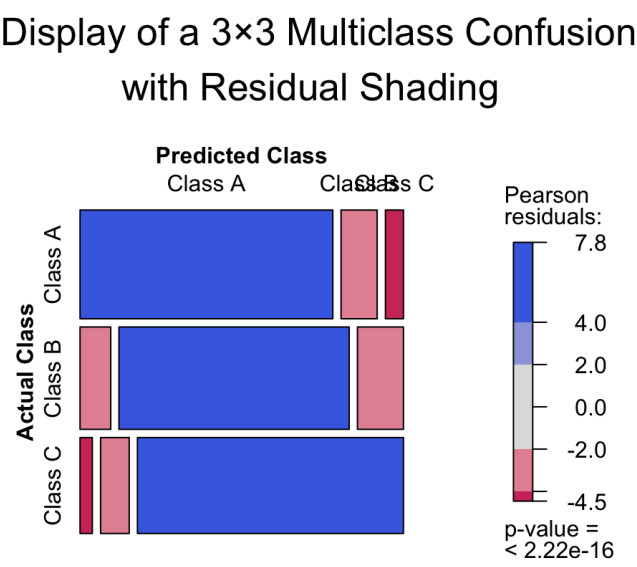

Figure 4. Mosaic display of a 3x3 multiclass confusion matrix with residual shading. Blue cells indicate higher-than-expected counts and red cells indicate lower-than expected counts under the independence model, revealing class-specific patterns of correct classification and systematic misclassification.

From a CDA standpoint, this decomposition aligns with the broader strategy of examining marginal and conditional structures within contingency tables. Rather than relying solely on global summaries, the analysis focuses on lower-dimensional margins and sub-tables that reveal specific patterns of association and misclassification (Agresti, 2013). This perspective is particularly valuable in multiclass settings, where the complexity of the full table can obscure important relationships between categories.

Another important consideration in multiclass confusion matrices is the heterogeneity of misclassification costs. In many applications, not all errors have equal consequences. For example, misclassifying a severe condition as mild may be more serious than the reverse. While standard confusion matrices treat all misclassifications equally, extensions such as cost-sensitive classification incorporate differential weights for different types of errors (He & Garcia, 2009). From the perspective of contingency tables, this corresponds to applying weights to cells, thereby modifying the interpretation of agreement and association.

Multiclass confusion matrices extend the basic ideas of binary classification into a more complex setting, where issues of imbalance, interpretability, and error structure become central.

While global measures such as accuracy provide limited insight, approaches based on decomposition into binary or pairwise comparisons offer a more informative and theoretically grounded analysis. By viewing these matrices through a CDA lens, it becomes possible to apply established principles for analyzing contingency tables to modern classification problems, therefore bridging classical statistical methods with contemporary applications in predictive modeling.

Visualizing Confusion Matrices and Classification Performance

Although confusion matrices are often presented as simple numerical tables, their interpretation is inherently visual. Usually, patterns are looked for: whether observations concentrate along the diagonal, whether errors cluster in particular off-diagonal cells, whether the marginal distributions reveal imbalance, and so forth. They concern the structure of association between observed and predicted classifications, which places confusion matrices directly within the graphical tradition of CDA (Friendly, 1994; Friendly, 2000).

A useful starting point is to recognize that the visual logic of a confusion matrix differs from the visual logic of many standard statistical plots. In scatterplots or line graphs, attention is often drawn to trends across continuous scales. In a confusion matrix, however, the analyst is reading a square table as a structured categorical display. The diagonal cells of a confusion matrix represent agreement between observed and predicted classifications, while the off-diagonal cells represent misclassification. However, a visually large diagonal cell does not necessarily indicate strong classification performance. It may simply reflect a large marginal total for a common class. This is why raw frequency displays can be misleading, especially when class distributions are imbalanced. A useful visualization should therefore show not only cell counts, but also the influence of margins, conditional proportions, or deviations from expected counts.

Mosaic displays are especially useful in this context because they represent cell frequencies through proportional areas, making the marginal structure of the table visible (Friendly, 1994). Figure 5 demonstrates this principle, showing how a mosaic display of a 2x2 confusion matrix makes both the marginal structure and deviations from independence visually explicit through proportional cell areas and residual shading. Figure 6 presents the same multiclass data as a conventional heatmap; while visually accessible, this format obscures the marginal structure of the table and provides no indication of which cells deviate from

independence. Applied to a confusion matrix, a mosaic plot can show whether performance is evenly distributed across classes or dominated by a majority category. This is more informative than a simple heatmap because it makes class imbalance visually explicit rather than leaving it hidden in the margins.

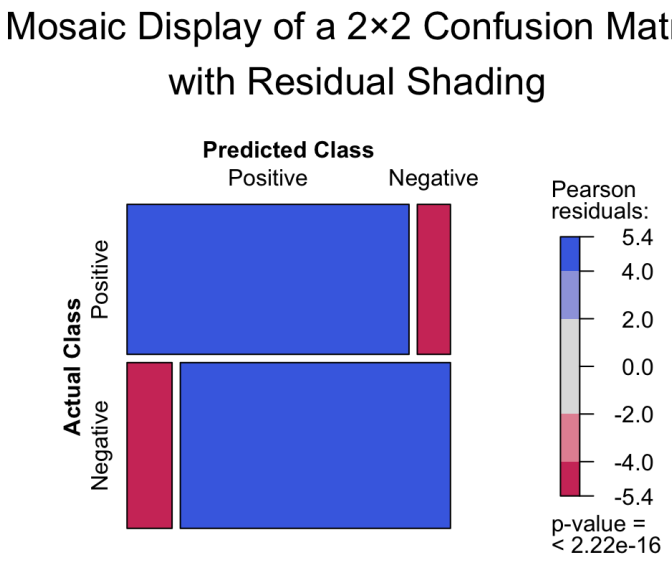

Figure 5. Mosaic display of a 2x2 confusion matrix with residual shading. Cell areas are proportional to observed frequencies; blue shading indicates higher-than-expected counts and red shading indicates lower-than expected counts under independence.

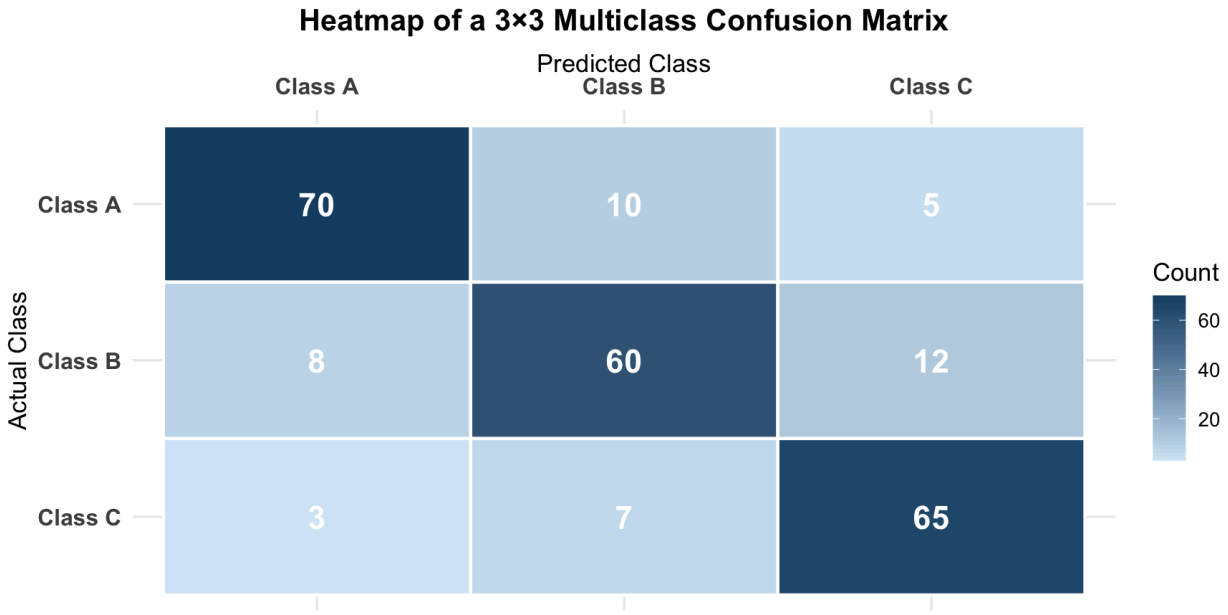

Figure 6. Heatmap of a 3x3 multiclass confusion matrix. Cell colour intensity reflects observed frequency, with darker shading indicating higher counts.

Residual-based shading extends this idea further by showing whether particular cells occur more or less often than expected under a baseline model, such as independence (Friendly, 1994; Zeileis, Meyer, & Hornik, 2007). This distinction is important because large cells are not always meaningful if their size is explained by marginal totals. Conversely, a smaller off-diagonal cell may be important if it occurs more frequently than expected. In this way, residual shading turns the confusion matrix from a descriptive summary into a visual diagnostic of association and misclassification structure.

Visual approaches are also useful for understanding diagnostic measures. Johnson's 2x2 diagram illustrates how quantities such as sensitivity, specificity, predictive values, and prevalence are tied to different margins of the same table (Johnson & Johnson, 2014). This is relevant because classification metrics are conditional proportions: sensitivity conditions on actual positives, specificity on actual negatives, and predictive values on predicted classifications. Visualizing the table helps clarify that these metrics answer different questions.

Visualization becomes even more important in multiclass confusion matrices. As the number of categories increases, the number of off-diagonal cells grows quickly, making it difficult to interpret errors using global summaries such as accuracy alone. Lovell and colleagues

(2021) argue that multinomial confusion matrices should be examined carefully because aggregate measures can obscure the class-pair errors that matter most, particularly under class imbalance. Visual displays help reveal whether errors are concentrated between particular categories, whether they are symmetric, and whether certain classes are systematically over- or under-predicted.

Overall, visualization is not simply a presentation tool; it is part of the analysis of confusion matrices as categorical data. Heatmaps reveal the concentration of counts, mosaic plots reveal the marginal structure, and residual shading reveals deviations from independence. Together, these graphical methods support the central argument of this paper: confusion matrices should be interpreted not only as classification performance summaries, but as structured contingency tables that reveal patterns of association, agreement, imbalance, and systematic misclassification.

Confusion Matrices as CDA Objects

The preceding sections have shown that confusion matrices are more than performance summaries. From a CDA perspective, they are square contingency tables that cross-classify two categorical variables: the observed and predicted class. This interpretation allows confusion matrices to be analyzed using the same ideas that apply to other contingency tables, including association, agreement, independence, and residual patterns (Agresti, 2013; Friendly & Meyer, 2016).

The independence model provides a useful baseline. If observed and predicted classifications are independent, then the classifier provides no meaningful information about the true class. In this case, expected cell counts are determined only by the row and column margins. A useful classifier should therefore produce a table that deviates from independence, especially through higher-than-expected counts along the diagonal. Standard CDA tools such as Pearson's chi-square statistic, likelihood-ratio chi-square tests, and residual analysis can be used to examine these departures (Agresti, 2013; Haberman, 1973; Friendly, 1994).

This perspective also connects confusion matrices to agreement analysis. In square tables, diagonal cells have special importance because they represent agreement between two classifications. In confusion matrices, the same diagonal structure represents correct classification. Measures such as Cohen's kappa summarize agreement beyond chance, while model-based approaches allow the researcher to examine whether agreement is uniform across

categories or concentrated in particular classes (Cohen, 1960; Tanner & Young, 1985; Agresti, 2013). This is important because a single measure of agreement can hide meaningful variation across the table, particularly when some categories are much more common than others.

Recent work on classification evaluation reinforces the importance of looking beyond global performance summaries. Lovell et al. (2021) emphasize that multinomial confusion matrices should be examined through class-specific and pairwise comparisons because aggregate measures can obscure important misclassification patterns, especially under class imbalance. Similarly, Thölke et al. (2023) demonstrate that accuracy can be misleading when class distributions are highly imbalanced, since performance may be driven mainly by correct classification of the majority class. More recent work has also argued that confusion-matrix-based measures should be interpreted with attention to uncertainty and the distribution of errors, rather than treated as fixed summaries of model performance (Lovell et al., 2023; Vanacore, Pellegrino, & Ciardiello, 2024).

In sum, treating confusion matrices as CDA objects strengthens their interpretation. Accuracy, sensitivity, specificity, and precision summarize selected aspects of the table, but they do not fully describe its structure. CDA methods allow the entire table to be examined, including its margins, diagonal concentration, off-diagonal errors, deviation from independence, and class-specific patterns of misclassification. This broader view supports the central argument of the paper: confusion matrices should be understood not only as tools for evaluating classifiers, but as structured categorical data that can be analyzed using the concepts and methods of CDA.

Conclusion

Confusion matrices are often used as practical tools for summarizing classification performance, but this paper has shown that they can also be understood as square contingency tables within CDA. This perspective connects familiar measures such as sensitivity, specificity, precision, and accuracy to broader CDA concepts, including association, independence, agreement, marginal structure, and residual patterns. Viewing confusion matrices in this way allows researchers to move beyond isolated performance metrics and examine the full structure of the relationship between observed and predicted classifications.

This framework is particularly useful when classification problems become more complex, such as in multiclass settings or when class distributions are highly imbalanced. In these cases, global measures like accuracy may conceal important patterns of misclassification,

while CDA tools and visual methods can reveal class-specific errors, diagonal agreement, and off-diagonal structure. Thus, confusion matrices should be interpreted not only as summaries of predictive success, but as structured categorical data objects that require careful statistical analysis.

Future work could further explore how loglinear models, agreement models, and residual-based visualizations can be used to analyze confusion matrices more formally. Additional research could also examine how uncertainty in confusion-matrix measures should be reported, especially in small samples or imbalanced datasets. As classification methods continue to expand across fields such as medicine, psychology, and machine learning, CDA offers a useful framework for making confusion matrices more interpretable, rigorous, and statistically meaningful.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. *BMJ, 308*(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *BMJ, 309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89*(425), 190–200. https://doi.org/10.1080/01621459.1994.10476460

Friendly, M. (2000). *Visualizing categorical data*. SAS Institute.

Friendly, M., & Meyer, D. (2016). *Discrete data analysis with R: Visualization and modeling techniques for categorical and count data*. CRC Press. https://doi.org/10.1201/b19022

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29*(1), 205–220. https://doi.org/10.2307/2529686

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Johnson, K. M., & Johnson, B. K. (2014). Visual presentation of statistical concepts in diagnostic testing: The 2 × 2 diagram. *AJR American Journal of Roentgenology, 203*(1), W14–W20. https://doi.org/10.2214/AJR.13.11954

Lovell, D., McCarron, B., Langfield, B., Tran, K., & Bradley, A. P. (2021). Taking the confusion out of multinomial confusion matrices and imbalanced classes. In Y. Xu, R. Wang, A. Lord, Y. L. Boo, R. Nayak, Y. Zhao, & G. Williams (Eds.), *Data mining: AusDM 2021* (Communications in Computer and Information Science, Vol. 1504, pp. 16–30). Springer. https://doi.org/10.1007/978-981-16-8531-6_2

Lovell, D., Miller, D., Capra, J., & Bradley, A. P. (2023). Never mind the metrics—what about the uncertainty? Visualising binary confusion matrix metric distributions to put performance in perspective. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 202, pp. 22702–22757). PMLR. https://proceedings.mlr.press/v202/lovell23a.html

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association, 80*(389), 175–180. https://doi.org/10.1080/01621459.1985.10477157

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage, 277*, Article 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Vanacore, A., Pellegrino, M. S., & Ciardiello, A. (2024). Fair evaluation of classifier predictive performance based on binary confusion matrix. *Computational Statistics, 39*, 363–383. https://doi.org/10.1007/s00180-022-01301-9

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics, 16*(3), 507–525. https://doi.org/10.1198/106186007X237856

AI Disclosure

I used ChatGPT as a support tool during the development of this paper. Specifically, I used it to help summarize and simplify certain parts of readings so I could better understand difficult concepts related to confusion matrices, agreement, class imbalance, and categorical data analysis. I also used AI for technical assistance with the R-generated figures, including troubleshooting code errors and refining the appearance of heatmaps and mosaic plots. All final writing, interpretation, source selection, figure inclusion, and revisions were reviewed and implemented by me.