European Research Studies Journal
Volume XXIX, Issue 2, 2026

pp. 345-356

Evaluation of Agreement-Related E-mail Classification Models with Unbalanced Classes

Submitted 03/04/26, 1st revision 25/04/26, 2nd revision 11/05/26, accepted 06/06/26

Marcin Hernes[1], Artur Rot[2], Ewa Walaszczyk[3], Janusz Tyburcy[4], Abigail Hańczyk[5]

Abstract:

Purpose: *The aim of the research is to evaluate the effectiveness of classification models of agreement-related emails with imbalanced classes, which allows for a more comprehensive assessment of their performance under severely imbalanced data and a better understanding of their behaviour in practical applications.*
Design/Methodology/Approach: *The following machine learning classification methods have been used: Complement Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine.*
Findings: *This research evaluated the effectiveness of classification models for agreement-related emails with imbalanced classes. Random Forest and Support Vector Machine achieve high values for both Accuracy and balanced Accuracy, demonstrating their strong classification performance.*
Practical Implications: *Random Forest and Support Vector Machine can be implemented in intelligent information systems for a mail dispatcher. Correspondence can be automatically routed to the person responsible for handling the inquiry. This speeds up the process and minimises the risk of an inquiry being overlooked or left unanswered.*
Originality/Value: *Despite a large body of research on email classification, there is still a lack of studies focused on specific applications, such as agreement document classification. In particular, it is rare to simultaneously examine different models and compare their performance using multiple metrics within a single real-world problem.*

Keywords: *Classification Methods, Intelligent information systems, business agreements, email dispatcher, machine learning.*

JEL codes: *C38, C88, M15.*

[1]*Assoc. Prof., Wroclaw University of Economics and Business, Department of Process Management, Poland, e-mail: marcin.hernes@ue.wroc.pl;*
[2]*Assoc. Prof., Wroclaw University of Economics and Business, Department of Information Systems, Poland, e-mail: artur.rot@ue.wroc.pl;*
[3]*Asst. Prof., Wroclaw University of Economics and Business, Department of Process Management, Poland, e-mail: ewa.walaszczyk@ue.wroc.pl;*
[4]*PROA Technology Sp. z o.o., Wrocław, Poland,*
e-mail: janusz.tyburcy@proatechnology.com;
[5]*The same as in 4, e-mail: abigail.hanczyk@proatechnology.com;*

Paper type: *Research article.*

Acknowledgements: *This research was funded by the National Centre for Research and Development (NCBR), Poland, under the Strategic Programme for Scientific Research and Development "Advanced Information, Telecommunication and Mechatronic Technologies – INFOSTRATEG", Project No. INFOSTRATEG-IV/0025/2022, entitled "Development and implementation of an intelligent correspondence dispatcher system for use in private companies and Public Utility Institutions (AI4POST)".*

1. Introduction

Intelligent information systems for a mail dispatcher are increasingly becoming the subject of scientific research and practical applications. Such a system identifies the need reported in an email and forwards it to the appropriate person in the company/office. Correspondence is automatically routed to the person responsible for handling the inquiry.

This speeds up the process and minimises the risk of an inquiry being overlooked or left unanswered. The customer or inquirer does not have to search for the appropriate department to address their inquiry. All they need to do is send the inquiry to a general address. The system also pre-prepares response forms, helping ensure high-quality service to customers/petitioners. The form includes instructions describing the procedure for the employee to fill in the information.

As a result, the completed document meets quality standards and minimises the risk of an incorrect decision. These systems are therefore designed to simplify, expedite, and improve the efficiency of handling matters related to serving the stakeholders of private companies or public institutions. This ultimately leads to an improvement in the services provided.

However, one of the main research problems in developing intelligent correspondence dispatching systems is email classification. Despite a large body of research, there is still a lack of studies focused on specific applications, such as agreement document classification. In particular, it is rare to simultaneously examine different models and compare their performance using multiple metrics within a single real-world problem.

This work addresses this gap by conducting a comparative analysis of selected classification models using several evaluation metrics. Therefore, the aim of the research is to evaluate the effectiveness of classification models of agreement-related emails with imbalanced classes. This enables a more comprehensive assessment of their performance under severely imbalanced data and a better understanding of their behaviour in practical applications.

The remainder of this paper is structured as follows. The next section presents related research in the considered field. Next, the research methodology is presented. The main sections of the paper present the results and discussion. The article ends with conclusions.

2. Related Research

The problem of classifying imbalanced data is one of the most important challenges in machine learning. In many practical applications, such as document classification or medical data analysis, some classes occur much more frequently than others. This imbalance complicates both model training and its reliable evaluation.

This issue is widely covered in the literature (Japkowicz and Stephen, 2002; He and Garcia, 2009; Branco *et al.,* 2016; Fernández *et al.,* 2018; Owusu-Adjei *et al.,* 2023; de la Cruz Huayanay *et al.,* 2024).

In such cases, a simple metric like Accuracy can be misleading, as a model can achieve high scores by focusing primarily on the dominant class. Therefore, other metrics, such as Precision, Recall, F1-score, or Balanced Accuracy, are used to reflect better how the model performs across all classes (He and Garcia, 2009).

Research also shows that Precision-Recall metrics are more useful than ROC metrics for highly imbalanced data (Saito and Rehmsmeier, 2015). Particularly for the legal documents and contracts analysed in this paper, where certain document types dominate, selecting appropriate metrics is crucial for correct model evaluation.

To improve model performance, various approaches are used to address data imbalance. These include increasing the number of minority-class examples (oversampling), decreasing the number of majority-class examples (undersampling), and approaches that account for different error costs. One popular method is SMOTE, which creates artificial examples for rare classes (Chawla *et al.,* 2002).

However, research indicates that the effectiveness of these methods depends on the type of data (Batista *et al.,* 2004). Particularly in legal document collections (as discussed in this paper), where minority classes may be severely underrepresented, choosing an appropriate data-balancing strategy can significantly impact results.

In text classification tasks, such as analysing contract documents, models such as Support Vector Machines, Random Forests, and Logistic Regression achieve good results. However, their performance can be compromised by uneven class distribution (Sebastiani, 2002; Kowsari *et al.,* 2019).

At the same time, for legal documents and contracts, which are characterised by highly repetitive structures and specialised language, models can achieve particularly high performance. It is also important to use appropriate validation

methods, such as stratified cross-validation, which preserve class proportions and obtain more reliable results (Kohavi, 1995).

Despite a large body of research, there is still a lack of studies focused on specific applications, such as contract document classification. In particular, it is rare to simultaneously examine different models and compare their performance using multiple metrics within a single real-world problem.

This work addresses this gap by conducting a comparative analysis of selected classification models using several evaluation metrics. This enables a more comprehensive assessment of their performance under severely imbalanced data and a better understanding of their behaviour in practical applications.

3. Research Methodology

3.1 Dataset Characteristics

The data used in the study contained contract content that is characterised by high predictability and repeatability, which allows models trained on such data to recognise and classify various contract elements effectively. The data showed significant variation in the number of samples across classes, which can affect the model's training balance. The number of files in each class is presented in Table 1.

The "Supply Agreement" class contains the vast majority of samples, with 42,604, indicating that it is the most common contract type in the dataset. The "Escrow Agreement" class comprises 1,549 samples, indicating a large number of these contracts, but significantly fewer than for supply agreements.

The "Lease Agreement" and "Service Agreement" classes contain 2,350 and 2,151 samples, respectively, which are typical for these contract types. The "Other Contract Type" class has 192 samples, indicating a moderate sample size. "Medical Services Agreement" has only 96 samples, making it one of the least represented classes. "Lending Agreement" has only 22 samples, which is by far the smallest class in the set.

Table 1. *Number of files in each class in the dataset.*

No.	Type of Agreement	Number of files
1.	Supply Agreement	42,604
2.	Escrow Agreement	1,549
3.	Lease Agreement	2,350
4.	Service Agreement	2,151
5.	Other Agreement Type	192
6.	Medical Services Agreement	96
7.	Lending Agreement	22

Source: *Authors' calculations.*

Models:
The study included the use of the following models:

- Complement Naive Bayes (CNB) – a variant of the multinomial Naive Bayes classifier designed to mitigate the poor performance of standard Naive Bayes under imbalanced class distributions. Instead of modelling the probability that an example belongs to a given class, CNB models the probability that an example does not belong to a given class, but belongs to its complement, which stabilises the decision boundary weights and reduces the bias toward majority classes (Rennie *et al.,* 2003; Anagaw and Chang, 2019; Seref and Bostanci, 2019).
- Logistic Regression – a basic statistical method and machine learning algorithm used to predict the probability of a specific event occurring. Unlike linear regression, which predicts continuous values, logistic regression is used for classification tasks, most often binary, where the result takes one of two values: 0 or 1. The result is interpreted as the probability of belonging to a given class. It is widely used due to its interpretability, low computational cost, and strong theoretical foundation in generalised linear models (Agresti, 2013; Iwagami *et al.,* 2024; Miyazaki *et al.,* 2024; Hu *et al.,* 2025).
- Random Forest – an ensemble learning method that constructs a set of decision trees and aggregates their predictions via majority voting (for classification purposes). Each tree is trained on a new data sample of the same size as the original set, created by repeated sampling with replacement, and uses a random subset of features at each split. This reduces overfitting and improves generalisation compared to a single tree. Random Forests are robust to noise, handle nonlinear relationships, and provide estimates of feature importance, making them a popular choice in machine learning research (Savargiv *et al.,* 2021; Salman *et al.,* 2024; Sun *et al.,* 2024).
- Support Vector Machine (SVM) – a margin-based classifier that seeks a separating hyperplane maximising the distance between the nearest training points of different classes, known as support vectors. SVMs are theoretically grounded in statistical learning theory and structural risk minimisation, and they often achieve competitive performance in high-dimensional spaces such as text and image classification (Burges, 1998; Cervantes *et al.,* 2020; Brunetti, 2023; Nordin *et al.,* 2023).

Metrics:
To evaluate the effectiveness of the model, the following metrics were used:

1. Precision: a measure of the quality of a classification model, defining how accurately the model classifies positive samples. It is the ratio of the number of correctly classified positive samples to the total number of samples the model classified as positive.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad (1)$$

2. Recall: a metric that measures how well a model detects positive cases. It is the ratio of the number of correctly classified positive samples to the total number of actually positive samples.

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \quad (2)$$

3. F1-score: the harmonic mean of precision and recall, a measure that balances these two metrics. It is particularly useful in cases where the data is unbalanced, as it accounts for both false positives and false negatives.

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

4. Accuracy: a general measure of a classification model's performance, measuring the percentage of all samples that are correctly classified. It is the ratio of the number of correct classifications to the total number of samples.

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Samples} \quad (4)$$

5. Balanced Accuracy: a metric that accounts for class imbalance in the data by calculating the average recall for each class. It is particularly useful when dealing with data in which one class is significantly more represented than others.

$$Balanced\ Accuracy = \frac{1}{2} \cdot (Recall_{class\ 1} + Recall_{class\ 2}) \quad (5)$$

To obtain reliable and repeatable results, we used the Repeated Stratified Cross-Validation method, which ensures that class proportions are preserved in both the training and test sets.

This procedure allows for more accurate model evaluation in the context of unevenly distributed data, where dominant classes can easily bias performance metrics such as Accuracy.

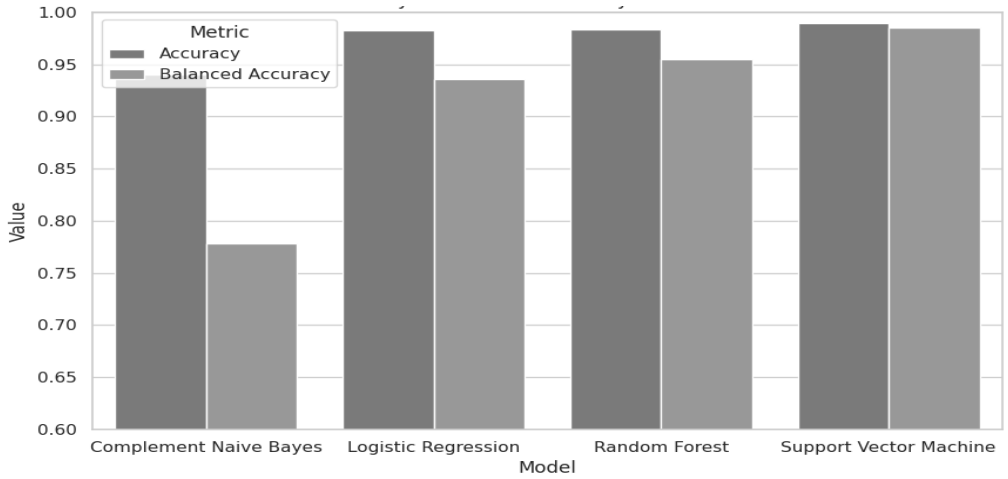
In particular, we included Balanced Accuracy, which considers performance on each class individually, a crucial metric in such cases. Analysis of the results enabled us to assess each model's ability to handle both dominant and minority classes, a crucial factor in practical applications.

3.2 Research Results and Discussion

Four machine learning models were used for contract classification: Complement Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine. Their Accuracy and Balanced Accuracy were compared. The resulting metrics are presented in Figure 1. The Complement Naive Bayes model achieved high Accuracy (93,99:), but its Balanced Accuracy (72,82%) was significantly lower, indicating class balancing issues.

The Logistic Regression model achieved very high scores on both metrics (98,29% and 93,60%, respectively), making it more versatile. Random Forest and Support Vector Machine achieved similar values for both Accuracy (98,32% and 98,97%, respectively) and Balanced Accuracy (95,48% and 98,53%, respectively), demonstrating their high performance in contract classification.

Figure 1. *Accuracy vs Balanced Accuracy Across Models*



Source: *Authors' own work.*

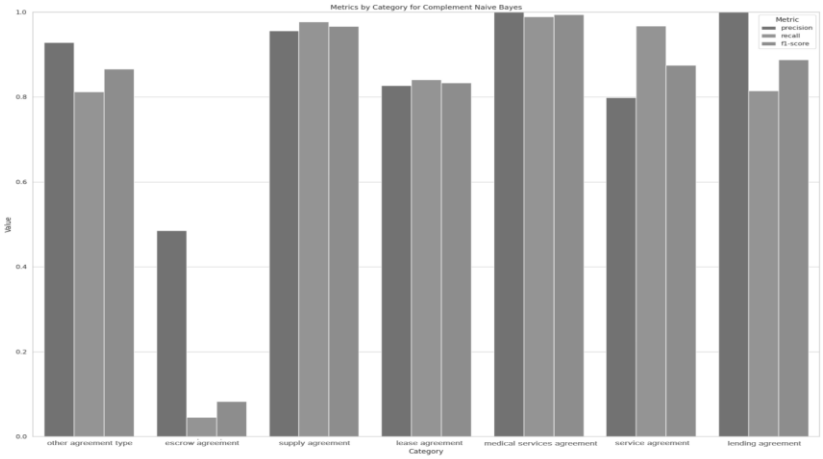
The results of each model for each class were also analysed. The Complementary Naive Bayes model achieved an Accuracy of 93.99% and a Balanced Accuracy of 77.82%, indicating good overall performance (Figure 2).

The best recognised class was "Supply Agreement" (F1-score: 96.64%), while the worst recognised class was "Deposit Agreement" (F1-score: 8.24%). Less represented classes, such as "Lending Agreement" (F1-score: 88.77%) and "Medical Services Agreement" (F1-score: 99.46%), performed well.

The Logistic Regression model achieved 98.29% Accuracy and 93.60% Balanced Accuracy, indicating very good performance (Figure 3). The best results were obtained for the "Loan Agreement" (F1-score: 100%) and "Delivery Agreement"

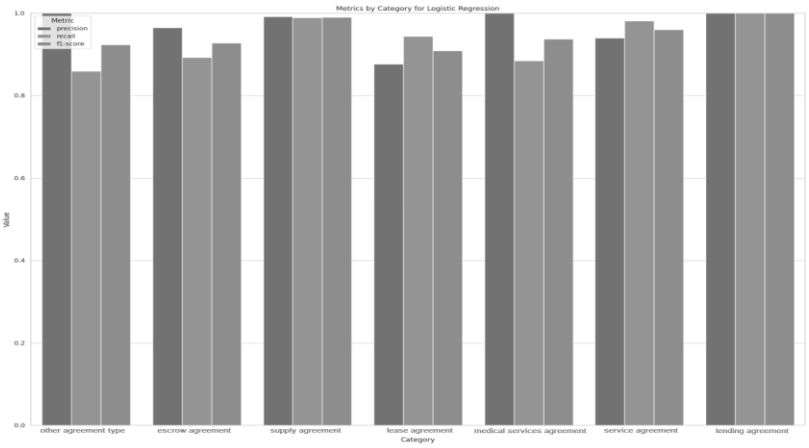
(F1-score: 99.05%) classes. The "Deposit Agreement" class was also well recognised (F1-score: 92.75%). The results for other classes, such as "Lease Agreement" (F1-score: 90.87%) and "Other Contract Type" (F1-score: 92.35%), were also high.

Figure 2. *Metrics by Category for Complementary Naive Bayes model.*



Source: *Authors' own work.*

Figure 3. *Metrics by Category for Logistics Regression model.*



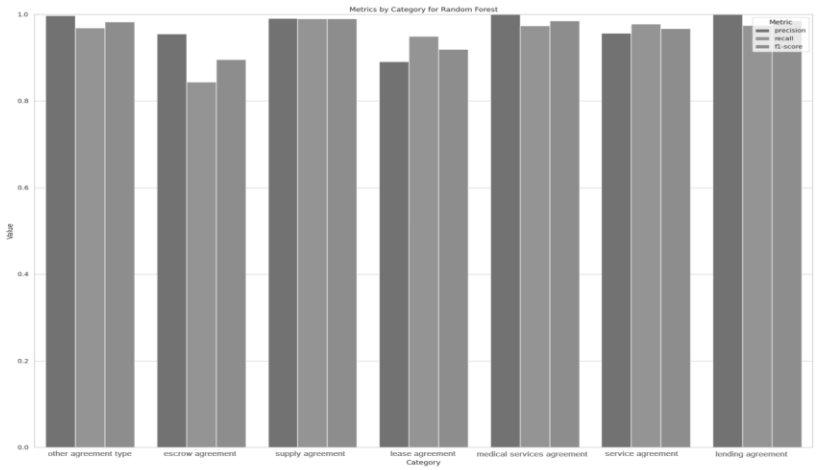
Source: *Authors' own work.*

The Random Forest model achieved an Accuracy of 98.32% and a Balanced Accuracy of 95.48%, which indicates very good performance (Figure 4). Classes

such as "Delivery Agreement" (F1-score: 99.06%) and "Lending Agreement" (F1-score: 98.57%) were recognised almost perfectly.

The class "Other Contract Type" also achieved a very high score (F1-score: 98.27%). The lowest results, but still high, were obtained for "Deposit Agreement" (F1-score: 89.59%).

Figure 4. *Metrics by Category for Random Forrest model.*



Source: *Authors' own work.*

The Support Vector Machine model achieved an Accuracy of 98.97% and a Balanced Accuracy of 98.53%, demonstrating its high performance (Figure 5). The "Medical Services Agreement" and "Loan Agreement" classes were classified perfectly (F1-score: 100%).

"Supply Agreement" also achieved an excellent result (F1-score: 99.43%), indicating excellent model performance for the most frequently encountered data.

Slightly lower, but still high, results were achieved for "Lease Agreement" (F1-score: 93.88%) and "Deposit Agreement" (F1-score: 96.45%). The weighted average results confirm very good classification quality across all classes.

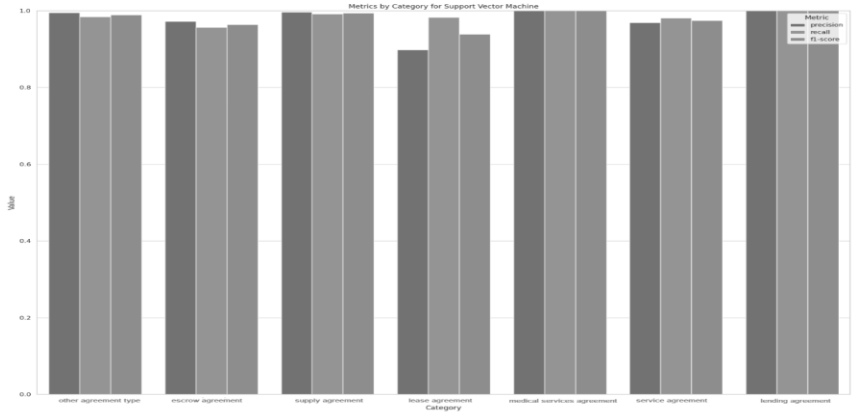
The data analysed in this study contained contract content characterised by high predictability and repeatability. Many of these documents had established structures, linguistic patterns, and repeated clauses, making them easy to analyse and process.

Thanks to this repeatability, models trained on such data could effectively recognise and classify various contract elements.

Thanks to these data characteristics, we achieved very high classification results. The repeatability and predictability of contract content enabled the models to learn effectively and correctly assign new documents to appropriate classes.

Techniques such as appropriate class weighting and the use of algorithms that handle unbalanced data enabled optimised results despite the uneven distribution of samples within individual classes.

Figure 5. *Metrics by Category for Support Vector Machine model.*



Source: *Authors' own work.*

4. Conclusions

This research evaluated the effectiveness of classification models for agreement-related emails with imbalanced classes. Random Forest and Support Vector Machine achieved high values for both Accuracy and Balanced Accuracy, demonstrating strong classification performance.

These models can be implemented in intelligent information systems for a mail dispatcher. The main limitation of the research is that it did not account for LLMs in the classification evaluation. Therefore, future research should include LLMs. Also, the data set should be expanded to include other types of agreements, such as contracts for different services and contracts for specific work.

References:

Agresti, A. 2013. Categorical Data Analysis. 3rd Edition, Wiley, Hoboken, New Jersey.

Anagaw, A., Chang, Y.L. 2019. A new Complement Naive Bayesian approach for biomedical data classification. J. Ambient Intell Human Comput, 10, 3889-3897. https://doi.org/10.1007/s12652-018-1160-1.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C. 2004. A study of the behavior of several

methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735.

Branco, P., Torgo, L., Ribeiro, R.P. 2016. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys, 49(2), 1-50. https://doi.org/10.1145/2907070.

Burges, C.J.C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, 121-167.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A. 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953.

de la Cruz Huayanay, A., Bazán, J.L., Russo, C.M. 2024. Performance of evaluation metrics for classification in imbalanced data. Computational Statistics, 40(3), 1447-1473. https://doi.org/10.1007/s00180-024-01539-5.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. 2018. Learning from imbalanced data sets. Springer. https://doi.org/10.1007/978-3-319-98074-4.

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.

Hu, Y., Zhang, X., Slavin, V., Belsti. Y., Tiruneh, S.A., Callander, E., Enticott, J. 2025. Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality. J. Med Internet Res, 27, e77721. https://www.jmir.org/2025/1/e77721.

Iwagami, M., Inokuchi, R., Kawakami, E., Yamada, T., Goto, A. 2024. Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study. PLOS Digital Health 3(8), e0000578. https://doi.org/10.1371/journal. pdig.0000578.

Japkowicz, N., Stephen, S. 2002. The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429-449. https://journals.sagepub.com/doi/10.3233/IDA-2002-6504.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1137-1143. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf.

Kowsari, K., Heidarysafa, M., Brown, D.E., Meimandi, K.J., Barnes, L.E. 2019. Text classification algorithms: A survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150.

Miyazaki, Y., Kawakami, M., Kondo, K. 2024. Logistic regression analysis and machine learning for predicting post-stroke gait independence: a retrospective study. Sci Rep 14, 21273. https://doi.org/ 10.1038/s41598-024-72206-4.

Nordin, N.I., Mustafa, W.A., Lola, M.S., Madi, E.N. 2023. Enhancing COVID-19 Classification Accuracy with a Hybrid SVM-LR Model. Bioengineering, 10(11), 1318. https://doi.org/10.3390/bioengineering10111318.

Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., Abdul-Salaam, G. 2023. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. PLOS Digital Health, 2(11), e0000290. https://doi.org/10.1371/journal.pdig.0000290.

Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003).

Saito, T., Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432.

Salman, H.A., Kalakech, A., Steiti, A. 2024. Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2(1), 69-79. https://doi.org/10.58496/BJML/2024/007.

Savargiv, M., Masoumi, B., Keyvanpour, M.R. 2021. A New Random Forest Algorithm Based on Learning Automata. Comput Intell Neurosci, 5572781. https://doi.org/10.1155/2021/5572781.

Sebastiani, F. 2002. Machine learning in automated text categorisation. ACM Computing Surveys, 34(1), 1-47. https://doi.org/10.1145/505282.505283.

Seref, B., Bostanci, E. 2019. Performance of Naive and Complement Naive Bayes Algorithms Based on Accuracy, Precision and Recall Performance Evaluation Criterions. International Journal of Computing Academic Research (IJCAR), 8(5), 75-92.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., Liang, X. 2024. An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237(B), 121549. https://doi.org/10.1016/j.eswa.2023.121549.