ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 6, Issue 1, April 2026, pp. 37-45

Comparative Machine Learning Methods for ICD-10 Diagnosis Classification

Deddy Rahmadi [a,1,*], Muhammad Solihin [b,2], Grandianus Seda Mada [c,3], Wakhid Fitri Albar [d,4], Sophia Carolina Shani [b,5]

[a] Department of Mathematics, Faculty of Science and Technology, Sunan Kalijaga State Islamic University, Jl. Marsda Adisucipto Yogyakarta 55281, Indonesia
[b] Department of Industrial Engineering, Faculty of Science and Technology, Sunan Kalijaga State Islamic University, Jl. Marsda Adisucipto Yogyakarta 55281, Indonesia
[c] Department of Mathematics, Faculty of Agriculture, Science and Health, University of Timor, Jl. Km 09 Sasi Kefamenanu 85613, Indonesia
[d] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Jl. Sekaran Semarang 50229, Indonesia
[1] deddy.rahmadi@uin.suka.ac.id *; [2] 22106060038@student.uin-suka.ac.id; [3] grandianusmada@unimor.ac.id;
[4] wakhid.albar@mail.unnes.ac.id; [5] 22106060039@student.uin-suka.ac.id
* Corresponding author

ARTICLE INFO

History
Submitted: October 9, 2025
Revised: November 19, 2025
Accepted: December 4, 2025

Keywords
ICD-10
Machine Learning
Random Forest
Decision Tree
SVM

ABSTRACT

The classification of disease diagnoses using the International Classification of Diseases (ICD-10) standard is essential for supporting clinical decision-making and administrative processes in healthcare systems. This study evaluated the performance of three machine learning algorithms, namely decision tree, random forest, and support vector machine (SVM), for ICD-10 diagnosis classification using 3,730 textual medical record entries collected from the Klinik Pratama UIN Sunan Kalijaga, Yogyakarta, Indonesia. The dataset exhibited significant class imbalance, which was addressed using the synthetic minority oversampling technique (SMOTE). The preprocessing procedures included text normalization and Term frequency-inverse document frequency (TF-IDF) vectorization, followed by model development with hyperparameter tuning through grid search cross validation. Model performance was assessed using accuracy, precision, recall, F1-score, confusion matrix, and five-fold cross validation. Random forest achieved the highest mean accuracy at 93.65%, followed by decision tree at 92.25% and SVM at 87.91%. These results indicate that ensemble-based approaches provide more reliable classification outcomes for imbalanced textual medical data. The findings are expected to support the development of semi-automated ICD-10 coding systems and improve the efficiency and accuracy of medical coding workflows.

1. Introduction

Disease diagnosis is one of the most critical stages in healthcare services because it directly influences decisions related to therapy, patient care, and hospital resource management [1]. The International Classification of Diseases, 10th revision (ICD-10), serves as a global standard adopted

p-ISSN 2798-253X
e-ISSN 2798-3153

in more than 100 countries for statistical reporting, epidemiological monitoring, and insurance claim processing [2], [3]. Despite its importance, the assignment of ICD-10 codes is still commonly performed manually by medical staff or medical record coders, making the process prone to inconsistency, human errors, and significant time consumption [4].

Previous comparative studies have assessed the performance of machine learning algorithms such as decision tree, random forest, and support vector machine (SVM) across various disease-related applications. Prior research reported that decision tree outperformed random forest and SVM in classifying COVID-19 symptoms, achieving a higher F1-score and a lower false positive rate [5]. Random forest obtained slightly better precision in diabetes prediction, although decision tree remained more computationally efficient [6]. Furthermore, decision tree achieved the highest accuracy on raw diabetes data, while k-nearest neighbor (KNN) performed better after min–max normalization [7]. These studies highlight that the effectiveness of machine learning models in medical diagnosis tasks is highly dependent on context, particularly with respect to preprocessing, class imbalance, and computational complexity [8].

However, there is limited research that directly compares random forest, decision tree, and SVM specifically for ICD-10 diagnosis classification in Indonesia, especially using a combination of textual diagnosis statements and structured patient metadata processed with TF-IDF [9]. This constitutes the main contribution of the present study. By integrating free-text diagnostic notes recorded by medical practitioners with patient-level variables such as gender, payment type, and month of visit, this research provides a more comprehensive evaluation of model performance in a real clinical environment. Therefore, this study aimed to conduct a comparative analysis of random forest, decision Tree, and SVM for ICD-10 diagnosis classification using clinical textual data and patient metadata. The findings are expected to support the development of semi-automated ICD-10 coding systems that improve accuracy, efficiency, and consistency in medical record workflows.

2. Method

2.1. Data Collection

This study was conducted at the Klinik Pratama UIN Sunan Kalijaga, Yogyakarta, Indonesia, using real-world patient visit data collected from January to December 2024. The dataset comprised 12 monthly digital records containing patient metadata such as the month of the visit, gender, payment type (BPJS or out-of-pocket), and diagnostic information coded using the International Classification of Diseases, 10th Revision (ICD-10) [10], along with textual diagnostic statements recorded by medical practitioners. All files were merged into a single master dataset following a comprehensive data cleaning procedure that involved removing duplicate entries, handling missing values, and resolving inconsistencies. Ethical approval for secondary data usage was obtained in accordance with data protection regulations. The final dataset consists of 3,730 diagnosis records covering 17 unique ICD-10 codes, with each instance corresponding to a single diagnostic statement associated with a patient visit.

The diagnosis statements were manually labeled by certified medical coders using the ICD-10 standard, ensuring high-quality annotations. To prepare the textual data for computational analysis, the diagnostic statements were then transformed into numerical feature representations using the TF-IDF method [9]. TF-IDF is a widely adopted technique for quantifying the importance of words within a document corpus, especially in healthcare natural language processing tasks, where identifying contextually meaningful terms is crucial. This transformation allows the extraction of relevant keywords and phrases that serve as predictors for ICD-10 code classification.

2.2. Data Preprocessing

Comprehensive data preprocessing was conducted to enhance the quality and consistency of the textual input. The preprocessing pipeline comprised text normalization, tokenization, vectorization, and class imbalance handling. In text normalization, all diagnostic texts were converted to lowercase, and extraneous elements such as punctuation marks, numeric values not related to codes, and stop

p-ISSN 2798-253X
e-ISSN 2798-3153

words were removed. This standardization helps reduce noise and improves the quality of extracted features. In tokenization, the cleaned text was then tokenized into individual words or tokens, providing a structured format for frequency analysis. Subsequently, in vectorization, tokenized words were transformed into numerical feature vectors using the TF-IDF technique. This step generates a weighted representation of each term based on its relevance within the entire corpus. In class imbalance handling, due to the naturally imbalanced distribution of ICD-10 codes in real hospital data, the synthetic minority over-sampling technique (SMOTE) was employed to balance minority classes by synthetically generating new instances. This step is critical to prevent model bias towards majority classes and to ensure fair performance across all diagnostic categories.

All preprocessing operations were implemented using Python libraries such as scikit-learn and imbalanced-learn, ensuring compatibility with subsequent model training steps. The class distribution in the original dataset was moderately imbalanced, with certain ICD-10 codes appearing significantly more frequently than others. To mitigate this, SMOTE was applied to the training set only, ensuring balanced class representation during model training. The dataset was split into 80% for training and 20% for testing using stratified sampling, preserving the class proportions in both subsets.

2.3. Model Building and Hyperparameter Tuning

The study compared three classification algorithms: decision tree classifier, random forest classifier, and SVM. These algorithms were chosen due to their proven effectiveness in medical text classification tasks. The decision tree classifier offers interpretable classification rules and clear decision paths, which are valuable for understanding how specific keywords influence diagnosis assignment [11], [12]. Meanwhile, the random forest classifier, as an ensemble method, combines multiple decision trees to improve prediction stability and generalization on large datasets [13], [14]. The SVM, known for its effectiveness in high-dimensional feature spaces, is well-suited for sparse textual data like TF-IDF vectors [15].

Comparing these contrasting approaches provides deeper insight into which paradigm is most suitable for ICD-10 text-based diagnosis classification. In addition, these three models were compared to address the lack of studies evaluating heterogeneous machine learning paradigms on ICD-10 diagnosis classification in Indonesia. By comparing a rule-based model, an ensemble method, and a margin-based classifier, this study identified which type of model could best handle the combination of textual diagnosis notes and patient metadata.

To achieve optimal model performance, hyperparameters were fine-tuned using grid search cross-validation [16]. Hyperparameter tuning was conducted to reduce model variance, avoid overfitting, and ensure robust generalization across different data partitions. This process is particularly important in medical classification tasks, where class imbalance and high-dimensional TF-IDF features require carefully optimized parameter settings. For the random forest classifier, the search space included variations in the number of trees, tree depth, minimum number of samples required to split a node, and minimum number of samples per leaf [17], [18]. The decision tree classifier was optimized using similar parameters, including tree depth and splitting criteria. For the SVM, the tuning process involved adjusting the regularization strength and evaluating performance using a linear kernel. In all models, class weighting strategies were applied to address the issue of class imbalance. The best-performing configurations were selected based on the highest mean cross-validation accuracy, ensuring robust generalization across different data partitions.

2.4. Evaluation Metrics

The performance of each classification model was evaluated using six quantitative metrics to ensure a comprehensive and reliable assessment of predictive quality [19]. Because the dataset contains imbalanced ICD-10 code classes, accuracy alone cannot fully describe model performance. Therefore, four additional metrics derived from the confusion matrix, namely precision, recall, and F1-score, were included to provide a more detailed evaluation of how well the model distinguished between majority and minority classes. The confusion matrix itself was also used to visualize which

ICD-10 codes were correctly or incorrectly classified. In addition, five-fold cross-validation was applied to ensure that the evaluation results are robust and generalizable across different data partitions [19], [20]. A summary of all evaluation metrics used in this study is presented in Table 1.

Table 1. Evaluation Metrics Used in This Study

No	Metric	Definition	Purpose in This Study
1	Accuracy	The ratio of correctly predicted observations to the total number of observations.	Measures overall correctness of the model's predictions.
2	Precision	The ratio of true positives to all predicted positives.	Assesses the model's capability to minimize false positives.
3	Recall (Sensitivity)	The ratio of true positives to all actual positives.	Evaluates how well the model identifies all relevant positive cases.
4	F1-Score	The harmonic mean of precision and recall.	Provides a balanced indicator when dealing with imbalanced classes.
5	Confusion Matrix	A tabular representation of actual versus predicted classifications for each class.	Visualizes which ICD-10 codes were correctly or incorrectly classified.
6	Cross-Validation	A statistical resampling technique for assessing model performance stability on multiple splits.	Ensures the evaluation results are robust and generalizable across various data subsets.

2.5. Implementation Environment

All computational procedures were conducted using the Google Colaboratory (Colab) cloud platform with Python 3. The implementation utilized several key libraries, including scikit-learn for model development and evaluation, pandas and NumPy for data preprocessing and numerical operations, and seaborn together with matplotlib for visualization. After hyperparameter optimization using grid search cross-validation, the optimized models were trained on the training dataset and evaluated on the hold-out test dataset to obtain the final performance metrics.

2.6. Research Method

The research followed a systematic workflow consisting of data collection, preprocessing, model development, hyperparameter optimization, and performance evaluation. The preprocessing stage included text preparation, feature extraction, and class imbalance handling, while machine learning models were trained and optimized using cross-validation techniques. The overall research workflow is presented in Fig. 1.

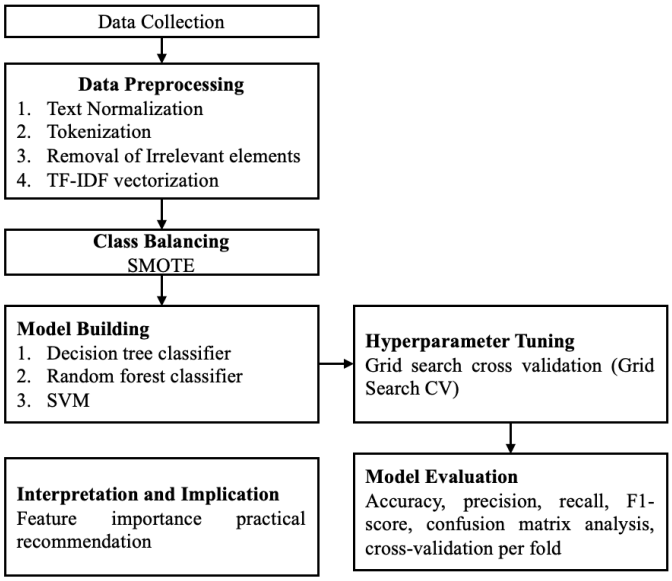


Fig. 1 Research flowchart.

The research concluded with the interpretation of results to identify strengths, limitations, and practical contributions for enhancing semi-automated ICD-10 coding processes in healthcare services.

3. Results and Discussion

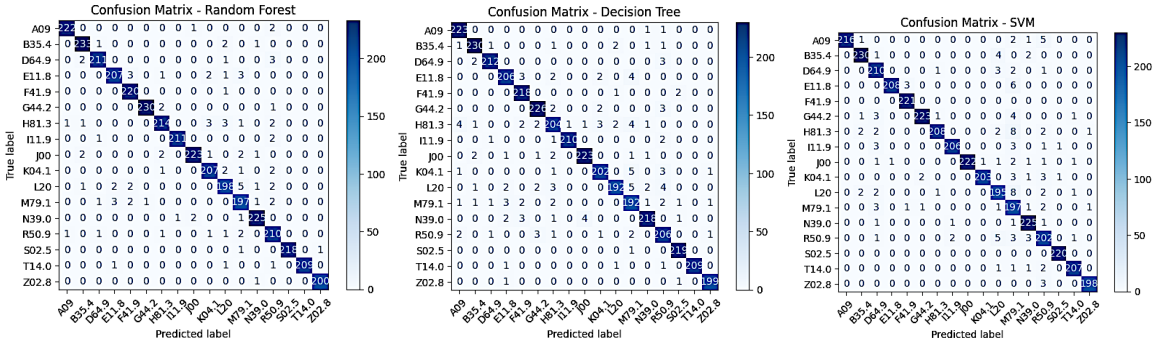
3.1. Model Performance Evaluation

To avoid biased evaluation and simulate real world deployment, the dataset used in this study, consisting of 6,978 real world ICD-10 diagnostic records containing both textual diagnosis notes and patient metadata, was split into 80% for training and 20% for testing. All hyperparameter tuning and cross validation were performed only on the training portion. The textual diagnosis statements were converted into numerical features using the TF-IDF method, which transforms text into weighted term frequency vectors suitable for machine learning classification. Model performance was then evaluated on a hold-out test set comprising 20% of the data (n = 746), which was not used during training or validation. On this independent test set the random forest model achieved an accuracy of 97%, while the decision tree and SVM models achieved accuracies of 96% and 97%, respectively.

Table 2. Precision, Recall, and F1-Score for Each ICD-10 Diagnosis Category

ICD-10 Code	Random Forest			Decision Tree			SVM		
	Precision	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***
A09	1.00	0.99	0.99	0.96	0.99	0.98	0.95	0.98	0.97
B35.4	0.97	0.98	0.98	0.97	0.97	0.97	0.98	0.98	0.98
D64.9	0.98	0.97	0.98	0.98	0.98	0.98	0.93	1.00	0.96
E11.8	0.98	0.95	0.97	0.95	0.95	0.95	0.99	0.95	0.97
F41.9	0.97	1.00	0.98	0.96	0.99	0.97	0.99	1.00	0.99
...	...	...	...	...	...	...	...	...	...
Z02.8	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.98	0.99
Overall Accuracy	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97

Table 2 summarizes the precision, recall, and F1-score for each diagnosis class. Random forest maintained consistently high values across all labels, with several classes achieving precision and recall above 0.95. For example, the class A09 shows a precision of 1.00 and recall of 0.99, indicating minimal misclassification for this category. On the other hand, the decision tree and SVM models achieved slightly lower recall for certain classes. For instance, the class M79.1 under the SVM shows a recall of 0.95, while random forest maintains comparable or higher values for the same label. Overall, these results suggested that the random forest model provided more stable classification performance across different ICD-10 categories, likely due to its ensemble. This stability is particularly important in medical text classification, where consistent performance across multiple diagnosis categories is essential for reliable clinical decision support. This advantage stems from the ensemble learning mechanism, which reduces variance and improves generalization on unseen data.



(a) (b) (c)

Fig. 2 Confusion matrix (a) random forest; (b) decision tree; (c) SVM, which were generated using Python 3 in the Google Colab environment.

Fig. 2 shows the confusion matrix for the random forest, decision tree, and SVM classifiers, respectively. The random forest matrix demonstrates clear diagonal dominance with very few off-diagonal elements, indicating that most predictions match the true labels. The decision tree and SVM matrices show similar patterns, but a closer look shows a slightly higher number of misclassified instances, especially for classes with overlapping keywords.

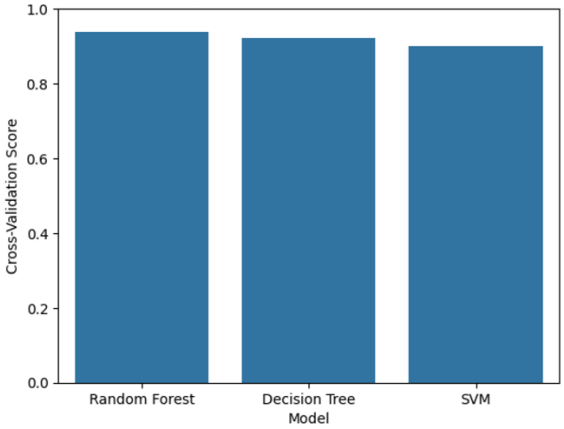


Fig. 3 Comparison of mean cross-validation scores for each model.

Fig. 3 presents the cross-validation scores for each model. Random forest achieved a mean cross-validation score of 0.9365, followed by the decision tree with 0.9225, and SVM with 0.8791. This result indicated that random forest not only performed well on training and test data but also maintained stable performance across multiple folds, which is important for practical implementation.

To ensure that the performance of each classification model was not biased by the specific partitioning of the dataset, a five-fold cross validation procedure was conducted for the tuned random forest, decision tree, and SVM models. Prior to this evaluation, hyperparameter tuning was performed using grid search cross validation, where key parameters for each model were optimized. For the random forest classifier, this included the number of trees, tree depth, minimum samples required to split a node, and minimum samples per leaf. The decision tree classifier was tuned using similar parameters, while the SVM tuning process involved adjusting the regularization strength and evaluating performance using a linear kernel.

In the subsequent cross validation stage, the dataset was divided into five equal subsets, sequentially using four-folds for training and one-fold for testing until all data had been used for validation exactly once. This validation strategy aimed to assess whether the models could maintain stable performance across different data splits and to detect any potential overfitting that might occur during the training phase.

Table 3. Cross-Validation Accuracy per Fold for Tuned Models

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Random Forest (Tuned)	0.9449	0.9374	0.9284	0.9381	0.9336	0.9365
Decision Tree (Tuned)	0.9396	0.9239	0.9187	0.9232	0.9068	0.9225
SVM (Tuned)	0.8934	0.8486	0.8762	0.865	0.912	0.8791

The cross-validation results per fold, shown in Table 3, represent the final performance of each model after hyperparameter tuning via grid search. Random forest exhibited relatively small variation between folds, with scores ranging from 0.9284 to 0.9449. The decision tree results ranged from 0.9068 to 0.9396, while SVM scores fluctuated more widely between 0.8486 and 0.9120. These findings indicated that the ensemble structure of random forest contributed to consistent predictive performance even when the training data was partitioned differently. The slightly greater score

variations observed in SVM suggested that linear kernels might be more sensitive to specific keyword distributions within the dataset.

To complement these cross-validation results, each tuned model was also evaluated on a hold-out test set comprising 20% of the dataset (n = 746), which was not involved in training or validation. The random forest classifier achieved a test accuracy of 97%, closely aligning with its cross-validation mean accuracy of 93.65%. Similarly, decision tree and SVM attained test accuracies of 96%, consistent with their respective cross-validation scores of 92.25% and 87.91%. This agreement between validation and test performance confirms that the models generalize well to unseen data and are not overfitted to the training partitions.

3.2. Feature Importance

In addition to performance metrics, this study also analyzed the contribution of individual keywords to classification accuracy. Feature importance was examined specifically for the random forest model because its ensemble structure provided more stable and reliable importance scores compared to a single decision tree, whose importance values tend to vary significantly with different training partitions. Moreover, SVM did not produce inherent feature importance measures in the same way, especially when applied to high dimensional TF-IDF representations. Therefore, random forest was selected as the most appropriate model for interpreting the relative influence of keywords, and Fig. 4 presents the top ranked features contributing to its classification decisions.

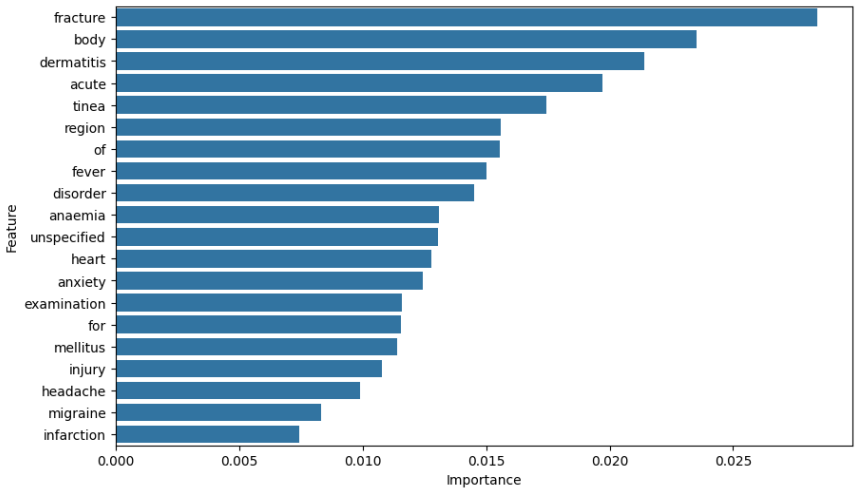


Fig. 4 Comparison of mean cross-validation scores for each model.

Keywords such as fracture, body, and dermatitis appear as dominant features, reflecting frequent diagnostic terms for orthopedic injuries, dermatological conditions, and related categories. The prominence of fracture aligns with the observation that trauma cases are among the most coded diagnoses in hospital settings. The appearance of terms like tinea and anaemia further illustrates the model's capacity to differentiate between infectious and chronic conditions.

This feature ranking agrees with the keyword weighting approach highlighted in [1] and supports the notion that relevant keyword extraction enhances ICD-10 mapping performance. Moreover, the diverse range of influential terms suggests that the random forest model does not overly rely on a narrow set of features, minimizing the risk of overfitting.

3.3. Practical Implications

The findings of this study demonstrated that the optimized random forest model achieved consistently high accuracy, precision, and recall across multiple ICD-10 diagnosis categories. This indicates strong potential for practical application in real-world clinical settings, particularly in supporting medical coders, health information managers, and hospital administration systems. By integrating this machine learning approach into the hospital information system (HIS), routine

diagnostic coding processes can be semi-automated, allowing human coders to focus on reviewing and validating model suggestions rather than performing repetitive manual coding. This can significantly reduce the time and administrative workload required to code large volumes of patient diagnoses, thereby improving the overall operational efficiency of medical record departments.

In addition, the high classification consistency across diverse diagnosis codes can help minimize coding errors, which is crucial for maintaining data quality in electronic health records and ensuring accurate reimbursement claims in health insurance systems. Accurate ICD-10 coding also supports better epidemiological reporting, resource allocation, and policy decision-making, especially in hospitals that process thousands of patient visits per month. However, practical deployment should consider integration with user-friendly interfaces that allow coders to view, accept, or override the model's suggestions in real-time. Moreover, to address the small misclassification rates identified for certain closely related diagnoses, future implementations could combine the random forest model with context-aware natural language processing (NLP) methods, such as word embeddings or transformer-based models, to capture subtler semantic nuances.

4. Conclusion

This study confirms that the use of machine learning algorithms, particularly random forest, decision tree, and SVM, offers substantial potential to automate ICD-10 diagnosis classification based on actual medical record data. Through a comparative approach, random forest demonstrated the highest overall accuracy and consistent cross validation performance, supported by strong precision and recall across multiple diagnostic categories. Decision tree also showed competitive results and provided clear interpretability, revealing the specific keywords and phrases most associated with each diagnostic label. In contrast, SVM achieved promising accuracy but showed greater sensitivity to the distribution of textual features, indicating that some ICD-10 categories contain more diverse or sparsely represented terms that require additional feature refinement.

Beyond model accuracy, the results also provide insight into the structure of the dataset itself. The feature importance analysis showed that certain medical terms, such as symptom descriptors and common diagnostic keywords, appeared more frequently and carried greater predictive weight for specific ICD-10 codes, while other categories with more variable or less frequent terminology remained harder to classify. These patterns suggest that the linguistic characteristics of diagnosis notes play a significant role in determining classification difficulty across ICD-10 groups.

The outcomes of this research emphasize that supervised learning methods can be practically integrated into semi-automated coding systems to assist medical coders, reduce manual workloads, and minimize potential human errors. The implementation pipeline, which combines patient metadata, TF-IDF text features, and class balancing through SMOTE, presents a reproducible framework that can be adapted to other healthcare institutions. In the future, this study is expected to support the development of more accurate and efficient medical record coding tools, contributing to higher data quality, more efficient hospital management, and better evidence-based health policies.

Acknowledgement

The author would like to thank Klinik Pratama UIN Sunan Kalijaga, Yogyakarta and other parties for their assistance and support in the sustainability of this research so that this paper can be completed.

References

[1] X. Zhan, M. Humbert-Droz, P. Mukherjee, and O. Gevaert, "Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases," *Patterns*, vol. 2, no. 7, Jul. 2021, doi: 10.1016/j.patter.2021.100289.

[2] World Health Organization, "International Statistical Classification of Diseases and Related Health Problems 10th Revision." 2015. [Online]. Available: https://icd.who.int/browse10/2015/en

[3] Z.A. Gafurov, "Classification, clinic and diagnosis of orbital factures," *Frontline Med. Sci. Pharm. J.*, vol. 02, no. 03, pp. 19–34, Mar. 2022, doi: 10.37547/medical-fmspj-02-03-03.

[4] R. Verma, A. Jain, and D. Ladsaria, "Automated extraction of ICD-10 diagnosis codes from clinical notes," 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:269313878.

[5] A. Wibowo, Indarti, and D. Laraswati, "Komparasi algoritma decision tree, random forest dan SVM untuk prognosis COVID-19," *IMTechno J. Ind. Manag. Technol.*, vol. 5, no. 2, pp. 10–15, Jul. 2024, doi: 10.31294/imtechno.v5i2.2868.

[6] A.F. Fadhlullah and T. Widiyaningtyas, "Comparative analysis of decision tree and random forest algorithms for diabetes prediction," *J. Teori Aplikasi Mat.*, vol. 8, no. 4, pp. 1121–1132, Oct. 2024, doi: 10.31764/jtam.v8i4.24388.

[7] S. Yulianty and M.K. Najib, "Comparing the accuracy of k-nearest neighbor (KNN), random forest, and decision tree methods in predicting diabetes," *Al-Aqlu J. Mat. Tek. Sains.*, vol. 3, no. 2, pp. 144–151, Jul. 2025. [Online]. Available: https://jurnal.yalamqa.com/index.php/aqlu

[8] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/nejmra1814259.

[9] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[10] B. Biswas, T.-H. Pham, and P. Zhang, "TransICD: Transformer based code-wise attention model for explainable ICD coding," 2021, *arXiv:2104.10652*.

[11] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., Boston, MA, USA: Springer, 2005, pp. 165–192.

[12] R. Khan, N. Ahmad, J. Ali, and I. Maqsood, "Random forests and decision trees," *Int. J. Comput. Sci.*, vol. 9, no. 5, pp. 272–278, Sep. 2012.

[13] Adeen and P. Sondhi, "Random forest based heart disease prediction," *Int. J. Sci. Res. (IJSR)*, vol. 10, no. 2, pp. 1669–1672, Feb. 2021, doi: 10.21275/sr21225214148.

[14] G. Louppe, "Understanding random forests: From theory to practice," Ph.D. dissertation, Dept. Electr. Eng. Comput. Sci, University of Liège, Liège, Belgium, 2014.

[15] K. Dharmarajan, K. Balasree, A.S. Arunachalam, and K. Abirmai, "Thyroid disease classification using decision tree and SVM," *Indian J. Public Health Res. Develop.*, vol. 11, no. 3, pp. 224–229, Mar. 2020, doi: 10.37506/IJPHRD.V11I3.822.

[16] M. Mohammadagha, "Hyperparameter optimization strategies for tree-based machine learning models prediction: A comparative study of AdaBoost, decision trees, and random forest," 2025. [Online]. Available: https://dx.doi.org/10.2139/ssrn.5226457.

[17] T. Kavzoglu, F. Bilucan, and A. Teke, "Comparison of support vector machines, random forest, and decision tree methods for classification of Sentinel-2A image using different band combinations," in *41st Asian Conf. Remote Sens.*, Deqing, China, 2020, pp. 2145–2152.

[18] P.W.S. Aji, Suprianto, and R. Dijaya, "Stroke disease prediction using random forest method," *KESATRIA J. Penerapan Sist. Inf. (Komp. Manaj.)*, vol. 4, no. 4, pp. 916–924, Oct. 2023, doi: 10.30645/kesatria.v4i4.242.g240.

[19] A. Zollanvari, "Model evaluation and selection," in *Machine Learning with Python*, A. Zollanvari, Ed., Switzerland: Springer Cham, 2023, pp. 237–281.

[20] T. Barwahwala, A. Mahajan, S. Mittal, and O. Reich, "Is Model Accuracy Enough? A Field Evaluation of a Machine Learning Model to Catch Bogus Firms," 2024. [Online]. Available: http://www.nber.org/papers/w32705

p-ISSN 2798-253X
e-ISSN 2798-3153