Beyond Equality Matching: Custom Loss functions for Semantics-Aware ICD-10 Coding

Monah Bou Hatoum [1][a], Jean Claude Charr [1][b], Alia Ghaddar [2,3][c], Christophe Guyeux [1][d], David Laiymani [1][e]

[1] *FEMTO-ST Institute, UMR 6174 CNRS, University of Franche-Comté, 90000 Belfort, France*

[2] *Department of Computer Science, the International University of Beirut, Beirut P.O. Box 146404, Lebanon*

[3] *Department of Computer Science, Lebanese International University, Beirut, Lebanon*

{*monah.bou_hatoum, jeanclaude.charr, christopheġuyeux, davidl̇aiymani*}*@univ-fcomte.fr, alia.ghaddar@liu.edu.lb*

Keywords: Deep Learning, Relevancy Comparison, Hierarchical Relationships, Semantic Similarity, Custom Loss Function, ICD-10 Coding, Cosine Similarity, Medical Coding Automation, Machine Learning in Healthcare

Abstract: **Background:** Accurate ICD-10 coding is vital for healthcare operations, yet manual processes are inefficient and error-prone. Machine learning offers automation potential but struggles with complex relationships between codes and clinical text. **Objective:** We propose a semantics-aware approach using custom loss functions to improve accuracy and clinical relevance in multi-label ICD-10 coding by leveraging cosine similarity to measure semantic relatedness between predicted and actual codes. **Methods:** Four custom loss functions (*True Label Cardinality Loss* (TLCL), *Predicted Label Cardinality Loss* (PLCL), *Balanced Harmonic Mean Loss* (BHML), and *Weighted Harmonic Mean Loss* (WHML)) were designed to capture hierarchical and semantic relationships. These were validated on a dataset of 9.57 million clinical notes from 24 medical specialties, using binary cross-entropy (BCE) loss as a baseline. **Results:** Our approach achieved a test micro-F1 score of 88.54%, surpassing the 74.64% baseline, with faster convergence and improved performance across specialties. **Conclusion:** Incorporating semantic similarity into the loss functions enhances ICD-10 code prediction, addressing clinical nuances and advancing machine learning in medical coding.

1 INTRODUCTION

The International Classification of Diseases (ICD) is a global standard for categorizing diseases, symptoms, and medical procedures, critical for healthcare operations such as billing, quality control, and clinical research (Otero Varela et al., 2021). Manual ICD-10 coding is inefficient, error-prone, and requires specialized knowledge (Mou et al., 2023; Zhou et al., 2020), driving the adoption of machine learning to automate this process (Esteva et al., 2019). However, existing models struggle with the complexity and ambiguity of medical data (Ghassemi et al., 2019).

A significant limitation of current models is their reliance on strict equality matching, penalizing predictions that deviate from exact matches (del Barrio et al., 2020; Long, 2021; Mittelstadt et al., 2023). This approach overlooks the clinical equivalence of certain codes (e.g., *Z01.8, Z01.9, Z48.8*) and fails to address hierarchical relationships in ICD-10, which are vital for accurate representation. Conversely, some codes (e.g., *P74.31, P74.32*) require strict specificity due to their distinct clinical implications (Hatoum et al., 2023). The ambiguity in clinical documentation further complicates this, as similar phrasing can correspond to different codes (Yu et al., 2023).

To overcome these challenges, we propose a relevancy-based approach leveraging vector representations of ICD-10 codes and cosine similarity to measure semantic relatedness. This method assigns partial credit for clinically valid predictions, enabling the model to handle nuanced relationships between codes effectively.

Our approach employs the Adam optimizer to address sparse gradients and class imbalance in large-scale datasets. We introduce four custom loss functions: *True Label Cardinality Loss (TLCL)*, *Predicted Label Cardinality Loss (PLCL)*, *Balanced Harmonic*

[a] https://orcid.org/0000-0002-0773-8409

[b] https://orcid.org/0000-0002-0807-4464

[c] https://orcid.org/0000-0003-1363-6174

[d] https://orcid.org/0000-0003-0195-4378

[e] https://orcid.org/0000-0003-2580-6660

Mean Loss (BHML), and *Weighted Harmonic Mean Loss (WHML)*. These optimize both accuracy and clinical relevance by capturing hierarchical and semantic relationships while minimizing penalties for clinically acceptable predictions.

Validated on a dataset of 9.57 million clinical notes spanning 24 specialties, our method demonstrated significant improvements in micro-F1 scores, outperforming traditional binary cross-entropy loss. By enhancing automated ICD-10 coding, this approach has the potential to improve healthcare efficiency, billing accuracy, and clinical decision-making.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 details the proposed loss functions, Section 4 presents the experimental setup and results, Section 5 discusses findings, and Section 6 concludes with future directions.

2 RELATED WORK

This section provides an overview of recent advancements in natural language processing (NLP) and their applications in healthcare, particularly in ICD-10 coding. First, the development of large language models (LLMs) and their potential in healthcare tasks is discussed. Second, the advantages of BERT-based models for clinical text classification are explored. Third, recent studies on custom loss functions in deep learning, with a focus on their applications in healthcare and medical coding, are reviewed. Finally, the latest developments in vector-based representations for ICD-10 codes, which directly relates to our proposed approach, are examined.

Recent advancements in natural language processing have been largely driven by the development of large language models (LLMs), such as *GPT-4* (Wu et al., 2023), *Claude 3* (Kurokawa et al., 2024), and *Gemini* (Mihalache et al., 2024). These models, trained on vast amounts of text data, have demonstrated remarkable capabilities in various tasks, including text generation, translation, and question-answering (Kumari and Pushphavati, 2022). The success of *LLMs* has sparked significant interest in their potential applications within the healthcare domain, particularly in tasks such as clinical text classification and ICD-10 code prediction (Al-Bashabsheh et al., 2023).

Building on these advancements, researchers have explored innovative ways to tailor *LLMs* to specific medical coding tasks. Although impressive, *LLMs* are computationally intensive and may not be feasible for healthcare organizations with limited on premise resources. Moreover, due to security and confidentiality purposes, sensitive medical data cannot be transferred to the available online LLMs. This has led to the adoption of more efficient models, particularly *BERT* (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which provides comparable effectiveness in classification tasks while requiring fewer computational resources (Mohammadi and Chapon, 2020; Zapata, 2023; Areshey and Mathkour, 2023; Grabner et al., 2022). *BERT*-based models, such as *ClinicalBERT* (Alsentzer et al., 2019), offer significant advantages due to their bidirectional nature, allowing them to capture the context and nuance within clinical text that is crucial for accurate ICD-10 code prediction. These characteristics make BERT-based models more practical and accessible for implementing automated coding solutions in healthcare settings (Areshey and Mathkour, 2023; Grabner et al., 2022).

In parallel, there has been growing interest in optimizing model training through the use of custom loss functions (Dinkel et al., 2019; Conley et al., 2021), particularly in complex domains like healthcare. Recent research has demonstrated that integrating contextualized loss functions into deep learning models can enhance their performance by optimizing the discovery of the relationships within the data, rather than focusing solely on exact matches. For instance, loss functions that prioritize relevant features and similarities have shown significant improvements in performance across various tasks, such as classification and image synthesis (Li et al., 2022; Kulkarni et al., 2024). In the context of ICD-10 coding, custom loss functions can play a crucial role in improving clinical relevance and prediction accuracy by moving beyond traditional equality-matching approaches (Yang et al., 2020; Boldini et al., 2022; Athanasiou and Maragoudakis, 2017). In our work, we build upon these insights to develop custom loss functions specifically tailored for ICD-10 coding, aiming to capture the complex relationships between clinical text and diagnostic codes.

Authors such as (Boldini et al., 2022) have demonstrated the effectiveness of custom loss functions for improving gradient boosting on imbalanced datasets. In their study, the introduction of an F-score loss and AUC loss optimized models for these metrics rather than traditional accuracy, resulting in significant improvements in F-score and AUC for imbalanced toxicity prediction tasks. Similarly, studies like (Wang et al., 2019) have highlighted the importance of custom loss functions in handling noisy labels, a common issue in medical datasets derived from electronic health records. The introduction of the Complementary Cross-Entropy (CCE) loss function in their work mitigated the effects of inaccurate data annotation, leading to improved classification accuracy.

Further, (Giyahchi et al., 2022) present an approach to fine-tuning pre-trained language models for intent detection, focusing on custom loss functions to capture nuanced relationships between user intents. Their study achieved remarkable accuracy rates, although limited to smoking cessation groups, showing the broader applicability of custom loss functions in healthcare-related NLP tasks.

Recently, advancements in ICD-10 code prediction from clinical notes have emphasized the importance of vector-based representations (Hatoum et al., 2024b). In (Hatoum et al., 2024a), the authors introduce *NNBSVR*, a method that generates semantic vector representations of ICD-10 codes. Their approach incorporates hierarchical relationships and contextual information, improving model accuracy by 12.73% on the test set compared to existing methods. Notably, their use of cosine similarity-based evaluation allows for a more nuanced assessment of model predictions, focusing on clinical relevance rather than strict exact matches. This research has direct implications for our work, as it aligns with our goal of integrating semantic vector representations with custom loss functions to optimize ICD-10 code predictions.

In summary, the literature reveals a clear trend toward more sophisticated *NLP* techniques in healthcare, particularly for ICD-10 coding. The progression from general language models to specialized approaches like *BERT* and the development of custom loss functions demonstrates the field's evolution. Our work builds on these advancements, particularly the vector-based representations introduced in (Hatoum et al., 2024a), by combining them with custom loss functions. This novel synergy, not fully explored in previous ICD-10 coding research, allows us to more effectively capture both the semantic relationships between codes and the nuanced information in clinical texts. Our approach aims to address the limitations of traditional equality-based methods, potentially leading to more accurate and clinically relevant predictions in ICD-10 coding.

3 CUSTOM LOSS FUNCTIONS FOR SEMANTIC-AWARE ICD-10 CODING

ICD-10 coding requires nuanced methods beyond traditional equality-based approaches, which fail to account for hierarchical and semantic relationships between codes. To address this, we propose four custom loss functions designed to capture these relationships while balancing specificity and flexibility. These loss functions integrate seamlessly with existing model architectures, enabling broad applicability across automated medical coding systems. Our focus is on optimizing performance through these loss functions without altering model structures.

3.1 Definitions

Let $X = \{x_i\}_{i=1}^{n}$ denote a dataset with n samples, each having a true label set $\mathcal{Y} = \{y_i\}_{i=1}^{n}$ and a predicted label set $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{n}$. Each y_i and $\hat{y}_i$ contains ICD-10 codes for the i-th sample. The complete set of unique ICD-10 codes is $\Lambda = \{\lambda_j\}_{j=1}^{m}$, where m is the total number of codes. Each code λ_j is represented as a d-dimensional vector $v_j = f(\lambda_j)$, where f maps codes to a semantic vector space. The cardinalities of the true and predicted label sets are $|y_i|$ and $|\hat{y}_i|$, respectively:

$$y_i = \{y_{ij}\}_{j=1}^{|y_i|}, \quad \hat{y}_i = \{\hat{y}_{ij}\}_{j=1}^{|\hat{y}_i|}$$

where y_{ij} and $\hat{y}_{ij}$ are individual labels in the true and predicted sets for sample x_i.

3.2 Formulation

Each true label y_{ij} and predicted label $\hat{y}_{ij}$ is mapped to vector representations v_{ij} and $\hat{v}_{ij}$. The semantic similarity between these vectors is computed using cosine similarity:

$$\cos(v_{ij}, \hat{v}_{ij}) = \frac{v_{ij}^{\top}\hat{v}_{ij}}{\|v_{ij}\|_2 \|\hat{v}_{ij}\|_2}$$

where $\|v_{ij}\|_2$ and $\|\hat{v}_{ij}\|_2$ are the L_2 norms of v_{ij} and $\hat{v}_{ij}$. A threshold $\tau \in [0,1]$ determines the strictness of similarity, with higher values requiring stronger similarity for a match. The binary indicator δ_{ij} determines if a true label y_{ij} matches a predicted label $\hat{y}_{ij}$:

$$\delta_{ij} = \begin{cases} 1, & \text{if } \cos(v_{ij}, \hat{v}_{ij}) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

Cosine similarity was chosen for its effectiveness in capturing semantic relationships in text classification tasks (Al-Anzi and AbuZeina, 2020). Its invariance to vector magnitude and computational efficiency make it particularly suitable for medical coding.

3.3 Custom Loss Functions

3.3.1 True Label Cardinality Loss (TLCL)

TLCL prioritizes recall, encouraging the model to predict all true labels. While this increases the capture of relevant codes, it may tolerate irrelevant predictions. The *TLCL* is defined as:

$$TLCL = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|y_i|}\sum_{j=1}^{|y_i|}(1-\delta_{ij})$$

where n is the number of samples, $|y_i|$ is the number of true labels for sample i, and δ_{ij} indicates a match. This ensures equal contribution from each true label.

3.3.2 Predicted Label Cardinality Loss (PLCL)

PLCL emphasizes precision, reducing irrelevant predictions by weighting each predicted label equally. It is defined as:

$$PLCL = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\hat{y}_i|}\sum_{j=1}^{|\hat{y}_i|}(1-\delta_{ij})$$

where $|\hat{y}_i|$ is the number of predicted labels for sample i. This formulation discourages false positives while being conservative in its predictions.

3.3.3 Balanced Harmonic Mean Loss (BHML)

BHML balances precision and recall by combining *TLCL* and *PLCL* using the harmonic mean:

$$BHML = \frac{2}{\frac{1}{TLCL}+\frac{1}{PLCL}}$$

The harmonic mean ensures that neither precision nor recall is disproportionately prioritized, creating a balanced optimization strategy.

3.3.4 Weighted Harmonic Mean Loss (WHML)

WHML generalizes *BHML* by introducing a weighting parameter $\alpha \in [0,1]$ to prioritize precision or recall as needed:

$$WHML = \frac{1}{\frac{\alpha}{TLCL}+\frac{1-\alpha}{PLCL}}$$

The parameter α controls the emphasis:

- $\alpha = 1$: Equivalent to *TLCL* (recall-focused).
- $\alpha = 0.5$: Equivalent to *BHML* (balanced).
- $\alpha = 0$: Equivalent to *PLCL* (precision-focused).

By adjusting α, *WHML* provides flexibility to adapt to various clinical scenarios, offering tailored trade-offs between precision and recall.

3.4 Loss Function Selection and Impact

The choice of loss function significantly affects model behavior:

- *TLCL* enhances recall, ensuring all true labels are captured.
- *PLCL* prioritizes precision, reducing irrelevant predictions.
- *BHML* and *WHML* provide balanced approaches, with *WHML* offering additional flexibility.

These loss functions allow for tailored optimization strategies, ensuring models better align with real-world ICD-10 coding requirements. By incorporating semantic similarity, clinically relevant but imperfect matches are appropriately handled, advancing automated medical coding systems' accuracy and relevance.

4 EXPERIMENTS AND RESULTS

This section evaluates the performance of our proposed custom loss functions for multi-label ICD-10 code prediction, demonstrating the value of leveraging vector code similarities and label cardinalities to improve clinical relevance.

4.1 Dataset

The dataset comprises 9.57 million clinical notes collected over three years from a private hospital. As shown in Figure 1, the data is imbalanced, with specialties like Internal Medicine (21.71%) and OB/GYN (12.06%) dominating, while others, such as Neurology, are underrepresented. This imbalance presents challenges for building models that generalize across both common and rare specialties.

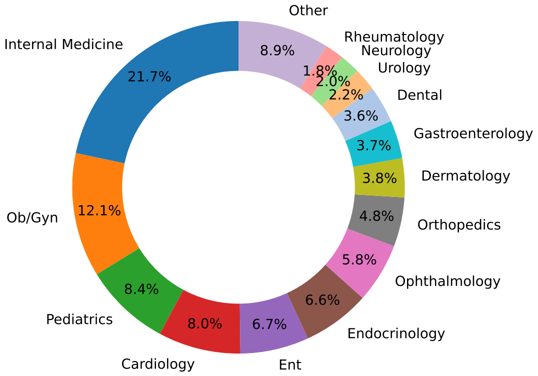

Figure 1: Dataset distribution across medical specialties.

Data preprocessing ensured consistency by standardizing medical terms, expanding abbreviations, and transforming values into categorical formats based on patient demographics, visit dates, and specialties (Hatoum et al., 2023). This preprocessing also addressed variability in writing styles across physicians from different countries, normalizing the clinical text for machine learning. The dataset contained 3,100 unique ICD-10 codes, and data handling adhered to stringent privacy regulations, ensuring confidentiality.

4.2 Setup

Clinical notes were tokenized using *BertTokenizer* (Devlin et al., 2018), and embeddings were generated using *ClinicalBERT* (Alsentzer et al., 2019). Labels were binarized into a 9.57M x 3,100 matrix using scikit-learn's MultiLabelBinarizer (Buitinck et al., 2013). The model architecture included *ClinicalBERT* as the embedding layer and a dense layer with sigmoid activations for multi-label predictions. The implementation was done using Keras (Chollet et al., 2015) and TensorFlow (Abadi et al., 2015). Hyperparameters are summarized in Table 1.

We compared baseline training with binary cross-entropy (BCE) (Mao et al., 2023) against the custom loss functions (*TLCL*, *PLCL*, *BHML*, and *WHML*). For *WHML*, we tested α values of 0.25 and 0.75 to explore precision-recall trade-offs. A cosine similarity threshold of $\tau = 0.76$ was identified as optimal through a grid search, achieving the best micro-F1 score (84.26%) on a sample dataset (Figure 2).

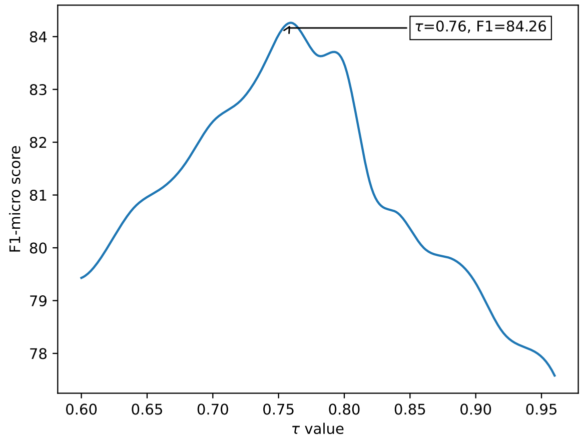

Figure 2: Micro-F1 scores for different cosine similarity ratios, with $\tau = 0.76$ yielding the best performance.

Parameter	**Value**
Embedding Layer	ClinicalBERT
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Early Stopping Patience	5
Number of Folds (Cross-Validation)	5
Number of Epochs (max)	50
Cosine Similarity Threshold	0.76

Table 1: Key hyperparameters for the experiments.

4.3 Results

4.3.1 Performance Comparison

Custom loss functions consistently outperformed the baseline BCE in both training and testing (Table 2). *WHML* with $\alpha = 0.75$ achieved the highest training F1-micro score (96.83%) and testing F1-micro score (88.54%). This superior performance highlights its ability to generalize across diverse ICD-10 codes, critical for multi-label classification tasks.

4.3.2 Specialty-Specific Performance

The *WHML* with $\alpha = 0.75$ achieved the highest F1-micro scores across most specialties, excelling in Pediatrics (97.51%), Ophthalmology (95.97%), and Dermatology (94.26%) (Table 3). *BHML* also performed well in precision-critical specialties like Dermatology and OB/GYN, while *TLCL* led in Cardiology (90.56%).

4.3.3 Performance on Challenging Codes

Custom loss functions excelled on challenging ICD-10 codes, improving accuracy for clinically ambiguous cases. For instance, *WHML* achieved a 10.19 percentage point improvement for ”K40.90 - Unilateral inguinal hernia” over BCE (Table 4), showcasing its ability to leverage semantic relationships.

These results highlight the advantages of relevancy-based loss functions in improving the accuracy and clinical relevance of automated ICD-10 coding systems, especially for complex and ambiguous cases.

5 DISCUSSION

Our results demonstrate the effectiveness of incorporating semantic similarity and hierarchical relationships into loss functions for ICD-10 code prediction. By moving beyond strict equality matching and considering clinical relevance, our approach significantly improves both accuracy and efficiency.

Training Results				**Testing Results**	
Experiment	**F1-micro**	**F1-Weighted**	**Epochs**	**F1-micro**	**F1-Weighted**
EM	83.75 ± 5.81e-03	84.31 ± 6.52e-03	22	74.64 ± 2.28e-03	72.01 ± 2.20e-03
TLCL	94.18 ± 2.56e-03	92.78 ± 3.02e-03	17	85.72 ± 1.89e-03	83.61 ± 2.11e-03
BHML	95.42 ± 2.32e-03	93.62 ± 3.51e-03	17	87.08 ± 1.95e-03	83.61 ± 2.34e-03
WHML $\alpha = 0.75$	96.83 ± 3.01e-03	94.71 ± 3.89e-03	17	88.54 ± 2.58e-03	86.92 ± 2.99e-03

Table 2: Training and testing results for baseline and custom loss functions (*TLCL*, *BHML*, *WHML*).

Specialty	**TLCL**	**BHML**	**WHML** $\alpha = 0.75$
Cardiology	90.56	91.64	**92.14**
Pediatrics	94.51	97.08	**97.51**
Emergency	74.26	76.34	**77.09**

Table 3: F1-micro scores for ICD-10 prediction across specialties.

5.1 Cost-effectiveness and Computational Complexity

The custom loss functions (*TLCL*, *PLCL*, *BHML*, and *WHML*) improve performance but introduce additional computational overhead due to cosine similarity calculations. For multi-label classification, the complexity for a single sample is $O(|y| \cdot d)$, where $|y|$ is the number of labels and d is the vector dimensionality, scaling to $O(n \cdot |y| \cdot d)$ for n samples. While more expensive than binary cross-entropy (*EM*), the faster convergence of *WHML* and *BHML* (compared to) offsets the higher per-iteration cost and enhanced the overall efficiency.

5.2 Limitations

Our study has some limitations. The dataset's regional focus (Saudi Arabia) and lack of rare ICD-10 codes (e.g., *W58 - Bitten by crocodile*) limit generalizability. Additionally, the dataset's imbalance, with overrepresented specialties, may reduce performance for underrepresented classes. Future work to address these issues could involve resampling or class weighting to improve robustness.

Moreover, this study focused on optimizing the loss functions rather than customizing the underlying model architecture or the optimizer. Future work could explore integrating customized optimizers and classifiers to further enhance the model's predictive power. This would allow us to tailor both the learning process and the architecture more closely to the needs of ICD-10 classification tasks, potentially unlocking even greater improvements.

5.3 Real-world Implementation and Future Work

One strength of our study is that the trained model has been implemented in a real-world hospital setting, where it is currently undergoing pilot testing. This provides valuable practical insights and demonstrates the feasibility of applying the proposed method in healthcare environments. Initial feedback from the pilot testing has been positive, though challenges have emerged, such as integrating the model into existing hospital workflows and ensuring compatibility with the hospital's electronic health record (EHR) systems. Additionally, the model's performance in handling ambiguous or incomplete clinical notes during real-time use is another area that requires further refinement.

Future research could improve data preprocessing through advanced semantic normalization and entity resolution, further enhancing model performance. Exploring specialty-specific configurations of the α parameter in *WHML* could optimize performance for different contexts. Additionally, customizing optimizers and classifiers may unlock greater accuracy and efficiency, tailoring the model to diverse healthcare needs.

By addressing these limitations and building on our findings, future work can further enhance automated ICD-10 coding, supporting more accurate healthcare decision-making and operational efficiency.

6 CONCLUSION

This study introduces semantics-aware loss functions for ICD-10 code prediction that incorporate clinical relevance and hierarchical relationships through vector representations. Our approach significantly out-

ICD-10-AM	Description	EM	TLCL	PLCL	BHML	WHML
J18.9	Pneumonia, unspecified	63.59	75.12	74.87	76.05	76.18
F41.9	Anxiety disorder, unspecified	53.17	56.94	56.32	57.21	57.29
M54.5	Low back pain	57.43	60.09	61.14	63.67	64.20
R10.4	Other and unspecified abdominal pain	50.38	58.40	57.78	59.17	59.64
G93.9	Disorder of brain, unspecified	49.08	50.21	50.14	51.02	50.83
K40.90	Unilateral or unspecified inguinal hernia without obstruction or gangrene, not specified as recurrent	59.86	67.22	65.47	69.13	70.05

Table 4: F1-scores for challenging ICD-10 codes across different loss functions.

performed traditional methods, achieving an 88.54

While cosine similarity calculations added computational overhead, faster convergence partially offset this cost. Pilot testing validates our approach's feasibility, though challenges remain in workflow integration and real-time processing. Future work will address ICD-10 code distribution variability, dataset imbalances, and specialty-specific optimization through refined WHML configurations and enhanced preprocessing techniques.

By improving automated medical coding accuracy and efficiency, our approach has the potential to streamline healthcare operations and support more informed clinical decision-making. With further refinements in model architecture and optimization strategies, these methods promise to advance both medical coding automation and healthcare analytics.

ACKNOWLEGEMENT

The authors would like to express their deepest gratitude to the medical coding department at Specialized Medical Center In Saudi Arabia. Their Unwavering support and dedication have been instrumental in the success of this research.

This work has been achieved in the frame of the EIPHI graduate school (Contract ”ANR-17-EURE-0002”).

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.

Al-Anzi, F. and AbuZeina, D. (2020). Enhanced latent semantic indexing using cosine similarity measures for medical application. http://dx.doi.org/10.34028/IAJIT/17/5/7.

Al-Bashabsheh, E., Alaiad, A., Al-Ayyoub, M., Beni-Yonis, O., Zitar, R. A., and Abualigah, L. (2023). Improving clinical documentation: automatic inference of icd-10 codes from patient notes using bert model. *The Journal of Supercomputing*, 79(11):12766–12790.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Areshey, A. and Mathkour, H. (2023). Transfer learning for sentiment classification using bidirectional encoder representations from transformers (bert) model. http://dx.doi.org/10.3390/s23115232.

Athanasiou, V. and Maragoudakis, M. (2017). A novel, gradient boosting framework for sentiment analysis in languages where nlp resources are not plentiful: A case study for modern greek. http://dx.doi.org/10.3390/a10010034.

Boldini, D., Friedrich, L., Kuhn, D., and Sieber, S. A. (2022). Tuning gradient boosting for imbalanced bioassay modelling with custom loss functions. http://dx.doi.org/10.1186/s13321-022-00657-w.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). Api design for machine learning software: experiences from the scikit-learn project. https://arxiv.org/abs/1309.0238.

Chollet, F. et al. (2015). Keras. https://keras.io.

Conley, T., Clair, J. S., and Kalita, J. (2021). Improving computer generated dialog with auxiliary loss functions and custom evaluation metrics. https://arxiv.org/abs/2106.02516.

del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2020). Review of mathematical frameworks for fairness in machine learning. https://arxiv.org/abs/2005.13755.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Dinkel, H., Wu, M., and Yu, K. (2019). Text-based depression detection on sparse data. https://arxiv.org/abs/1904.05154.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. http://dx.doi.org/10.1038/s41591-018-0316-z.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2019). A review of challenges and opportunities in machine learning for health. https://arxiv.org/abs/1806.00388.

Giyahchi, T., Singh, S., Harris, I., and Pechmann, C. (2022). Customized training of pretrained language models to detect post intents in online health support groups. http://dx.doi.org/10.1007/978-3-031-14771-5_5.

Grabner, C., Safont-Andreu, A., Burmer, C., and Schekotihin, K. (2022). A bert-based report classification for semiconductor failure analysis. http://dx.doi.org/10.31399/asm.cp.istfa2022p0028.

Hatoum, M., Charr, J.-C., Guyeux, C., Laiymani, D., and Ghaddar, A. (2023). Emte: An enhanced medical terms extractor using pattern matching rules. http://dx.doi.org/10.5220/0011717300003393.

Hatoum, M. B., Charr, J. C., Ghaddar, A., Guyeux, C., and Laiymani, D. (2024a). Nnbsvr: Neural network-based semantic vector representations of icd-10 codes.

Hatoum, M. B., Charr, J. C., Ghaddar, A., Guyeux, C., and Laiymani, D. (2024b). Utp: A unified term presentation tool for clinical textual data using pattern-matching rules and dictionary-based ontologies. http://dx.doi.org/10.1007/978-3-031-55326-4_17.

Kulkarni, D., Ghosh, A., Girdhari, A., Liu, S., Vance, L. A., Unruh, M., and Sarkar, J. (2024). Enhancing pre-trained contextual embeddings with triplet loss as an effective fine-tuning method for extracting clinical features from electronic health record derived mental health clinical notes. http://dx.doi.org/10.1016/j.nlp.2023.100045.

Kumari, S. and Pushphavati, T. (2022). Question answering and text generation using bert and gpt-2 model. In *Computational Methods and Data Engineering: Proceedings of ICCMDE 2021*, pages 93–110. Springer.

Kurokawa, R., Ohizumi, Y., Kanzawa, J., Kurokawa, M., Kiguchi, T., Gonoi, W., and Abe, O. (2024). Diagnostic performance of claude 3 from patient history and key images in diagnosis please cases. https://www.medrxiv.org/content/early/2024/04/14/2024.04.11.24305622.

Li, Z., Huang, X., Zhang, Z., Liu, L., Wang, F., Li, S., Gao, S., and Xia, J. (2022). Synthesis of magnetic resonance images from computed tomography data using convolutional neural network with contextual loss function. http://dx.doi.org/10.21037/qims-21-846.

Long, R. (2021). Fairness in machine learning: Against false positive rate equality as a measure of fairness. http://dx.doi.org/10.1163/17455243-20213439.

Mao, A., Mohri, M., and Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications.

Mihalache, A., Grad, J., Patil, N. S., Huang, R. S., Popovic, M. M., Mallipatna, A., Kertes, P. J., and Muni, R. H. (2024). Google gemini and bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. *Eye*.

Mittelstadt, B., Wachter, S., and Russell, C. (2023). The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. https://arxiv.org/abs/2302.02404.

Mohammadi, S. and Chapon, M. (2020). Investigating the performance of fine-tuned text classification models based-on bert. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1252–1257. IEEE.

Mou, C., Ye, X., Wu, J., and Dai, W. (2023). Automated icd coding based on neural machine translation. http://dx.doi.org/10.1109/ICCCBDA56900.2023.10154772.

Otero Varela, L., Doktorchik, C., Wiebe, N., Quan, H., and Eastwood, C. (2021). Exploring the differences in icd and hospital morbidity data collection features across countries: an international survey. http://dx.doi.org/10.1186/s12913-021-06302-w.

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., and Liu, H. (2019). A clinical text classification paradigm using weak supervision and deep representation. http://dx.doi.org/10.1186/s12911-018-0723-6.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., and Tang, Y. (2023). A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Yang, X., Bian, J., and Wu, Y. (2020). Customize deep learning-based de-identification systems using local clinical notes - a study of sample size. https://www.medrxiv.org/content/early/2020/10/26/2020.08.09.20171231.

Yu, Y., Qiu, T., Duan, J., and Wang, J. (2023). Multigranularity label prediction model for automatic international classification of diseases coding in clinical text. http://dx.doi.org/10.1089/cmb.2023.0096.

Zapata, L. (2023). Recommendation of text properties for short texts with the use of machine learning : A comparative study of state-of-the-art techniques including bert and gpt-2. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS).

Zhou, L., Cheng, C., Ou, D., and Huang, H. (2020). Construction of a semi-automatic icd-10 coding system. http://dx.doi.org/10.1186/s12911-020-1085-4.