1. Literature Review

The automation of medical coding using natural language processing (NLP) has been an active area of research for several decades given its potential to reduce costs, improve efficiency and data quality in healthcare systems. Early work in this field focused on rule-based and dictionary-based approaches that relied on handcrafted rules and lexicons to extract coded concepts from clinical text. However, these methods suffered from limited generalizability and the high overhead of manually curating and maintaining the rules.

In recent years, data-driven approaches based on machine learning have gained significant traction, fuelled by the availability of large annotated clinical corpora and advances in deep learning. In particular transfer learning techniques involving pre-trained language models have emerged as a powerful paradigm for adapting general-purpose representations to specialized domains like clinical NLP.

This literature review covers key research publications and state-of-the-art methods related to developing NLP systems for automated medical coding and clinical entity extraction, which is a crucial first step in the medical coding pipeline.

Deep Learning for ICD Coding

In a study using Pre-trained word embedding for medical coding Patel et al. (2017) proposes a method to adapt generic pre-trained word embeddings by incorporating information from a target domain and task. Word embeddings are dense vector representations of words that capture semantic and syntactic relationships. They have become a crucial component in modern natural language processing (NLP) tasks like part-of-speech tagging, named entity recognition, sentiment analysis, and others. The paper proposes a method to adapt pre-trained word embeddings by adding task-specific information to improve their performance. The study specifically uses a medical coding dataset and the hierarchy of ICD-10 medical codes to fine-tune widely used pre-trained embeddings like Word2Vec Mikolov et al. (2013). It adapts the continuous bag-of-words (CBOW) algorithm to treat the medical terms as context words and codes as target words during training. They evaluate an automated review of medical coding using 5 different pre-trained word embeddings - Google's and 4 medical domain-specific ones. Both original and adapted (modified) versions of embeddings are used with a logistic regression classifier. The adapted word embeddings consistently outperform the original embeddings across all 5 pre-trained sets. They obtain around 1% improvement in the F-score for the medical coding review task using the adapted embeddings. The paper demonstrates that adding domain knowledge can make generic pre-trained embeddings more suitable for specialized applications like medical coding. Some limitations were identified in the study. The paper only evaluates the modified word embeddings on a private medical claims dataset, which may limit the generalizability of the results. The paper mentions that adding extra information to pre-trained word embeddings is possible and beneficial, but it does not provide a comprehensive analysis of the potential limitations or drawbacks of this approach. The focus is only on adding information from a medical coding dataset and the first level of the ICD-10 code hierarchy. It does not explore the potential benefits or limitations of adding information from other domain-specific datasets or higher levels of the code hierarchy. There is no discussion of the potential impact of using different pre-trained word embeddings on the performance of the modified word embeddings. It only evaluates the approach on five different pre-trained word embeddings without comparing their performance to other state-of-the-art methods or baselines.

A neural network architecture proposed by Xie and Xing (2018) is capable of converting free-text diagnosis descriptions into relevant ICD codes. The architecture comprises several components, including a tree-of-sequences LSTM that encodes each code description using a sequential LSTM. These representations are then composed hierarchically with a bidirectional tree LSTM that follows the code tree structure, enabling the model to capture both the meaning of individual codes and their hierarchical relationships. Additionally, an adversarial learning component is included to make the encoded diagnosis and code descriptions indistinguishable in terms of writing style, mitigating the mismatch between their language domains. Furthermore, the model imposes isotonic constraints to ensure that the predicted probability scores of the codes adhere to the provided importance ordering during training. Finally, the model employs an attentional mechanism that computes an importance weight for each diagnosis description when predicting a particular code, allowing for flexible mappings. The paper presents a novel neural architecture tailored to the unique challenges of automated ICD coding, leveraging structured modelling, adversarial training, constraints, and attentional mechanisms. The paper mentions two major limitations of this study: It does not perform well on infrequent ICD codes. It is less capable of dealing with abbreviations in the diagnosis descriptions. The paper highlights two main limitations of the study. Firstly, it does not perform well on infrequent ICD codes. Secondly, it is less effective in dealing with abbreviations in the diagnosis descriptions. The paper proposes that future research could overcome these limitations by exploring diversity-promoting regularization to enhance the performance on infrequent codes and utilizing an external knowledge base to map medical abbreviations to their complete names.

The study by J. Huang et al. (2019) evaluates the performance of deep learning methods to automatically assign ICD-9 medical codes to clinical notes, using the MIMIC-III dataset. The study compares traditional machine learning algorithms, such as Logistic Regression and Random Forests, with state-of-the-art deep learning models, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTMs), and Gated Recurrent Units (GRUs).

The deep learning models significantly outperformed the traditional machine learning baselines for predicting the top 10 ICD-9 codes and categories, with the GRU model achieving the best F1 score of 0.6957 and 0.7233 respectively. However, for predicting the top 50 ICD-9 codes and categories, the results were mixed, with the traditional Logistic Regression model performing better. The authors suggest that this may be due to data imbalance issues for the less common codes.

The paper provides a thorough empirical evaluation and comparison of different deep-learning architectures and feature representations for this task. The authors also make the code and evaluation tools publicly available as a baseline for future research.

However, the study has a few limitations and gaps. The models did not perform as well for predicting the top 50 codes/categories as compared to the top 10. The authors suggest addressing this by increasing the sample size for the less common codes or using techniques like transfer learning. Additionally, the authors suggest exploring newer architectures like

Temporal Convolutional Networks and alternative text representations beyond just word2vec, such as sentence/paragraph embeddings, for further improvements.

A general and multilingual approach to automatically map diagnostic texts to ICD-10 codes using interpretable deep learning was presented in a study by Atutxa et al. (2019). They approached this as a sequence-to-sequence task and tested different neural network architectures, including RNN, CNN, and Transformer models. Their proposed approach outperformed previous state-of-the-art results on multilingual ICD-10 coding tasks in French, Hungarian, and Italian datasets. The system achieved F-measures of 0.838, 0.963, and 0.952 respectively, indicating high accuracy. Moreover, the system provides interpretable results by aligning each predicted ICD-10 code with the relevant parts of the input diagnostic text. This feature helps medical experts understand the reasoning behind the coding decisions.

The research analyses the impact of different neural architectures and the size of the training data. The authors found that Transformer-based models perform better on smaller datasets, while RNN and CNN-based models are more effective on larger datasets. The error analysis conducted by the authors identifies common issues like abbreviations, spelling errors, and alignment problems when dealing with multiple diagnostic terms in a single input. This analysis provides insights for future improvements.

However, the study also highlights a few gaps and limitations. For instance, when a single document line contains multiple diagnostic terms, there is no guaranteed 1-to-1 alignment between the terms and the assigned ICD-10 codes, which can cause alignment problems for some approaches. Additionally, the use of non-standard language and lexical variability in the diagnostic terms poses challenges, as the system needs to handle misspellings, abbreviations, and other deviations from standard terminology. While the neural network approaches generally performed well, the paper notes that the relative performance of different architectures (Transformer, CNN, RNN) can depend on the size of the training dataset. Larger datasets may favour CNN and RNN-based models over Transformer models.

The paper by Du et al. (2019) proposes a new deep learning framework called ML-Net for the multi-label classification of biomedical texts. It combines a label prediction network with an automated label count prediction mechanism to provide an optimal set of labels, leveraging both the predicted confidence score of each label and the deep contextual information in the target document. The document encoding network uses ELMo for contextualized word embeddings and a bidirectional RNN with attention to represent the document. ML-Net is evaluated on three multi-label text classification tasks in biomedical literature and clinical notes and is shown to outperform state-of-the-art machine learning and deep learning baseline models. ML-Net achieves the best F-scores on the biomedical literature tasks compared to baselines. However, for the diagnosis code assignment task with a large label space, thresholding-based methods perform better than ML-Net's label count prediction. ML-Net's label count prediction works well when labels are independent but struggles with hierarchical label structures.

There are some gaps and limitations identified in this work. Due to computation limitations, the authors could only process the first 1500 tokens of each clinical note for the diagnosis code assignment task. A more thorough hyperparameter tuning may not have been possible due to compute constraints. The proposed label count prediction network in ML-Net takes only the document vector as input and does not model the hierarchical relations between labels. For the tasks with hierarchical label structures like the diagnosis code assignment, the label count prediction did not perform as well as the thresholding approaches. The authors suggest that more advanced language representation models like BERT could potentially improve the document encoding component of ML-Net. The authors also note that the imbalanced data distribution, with some labels being much more frequent, can lead to ML-Net making more errors on predicting those common labels.

An unsupervised method for assigning ICD codes to clinical notes was proposed in a study by Kumar et al. (2022). The study uses a novel concept called Word Mover's Similarity (WMS) to compute the similarity between a clinical note diagnosis, description and ICD codes. This allows assigning relevant ICD codes to the note without requiring labelled training data. To speed up the computation, the authors introduce Relaxed Word Mover's Similarity (RWMS), which is a relaxed version of WMS that can be computed more efficiently. The method also takes into account the hierarchical structure of the ICD code tree to ensure the assigned codes are consistent. It performs a breadth-first search traversal of the code tree. The paper also introduces some rules, like handling negation words and n-grams, to further improve the performance of the unsupervised assignment. Experiments on the MIMIC-III dataset show that the proposed Word Mover's Similarity method outperforms other unsupervised baselines in terms of sensitivity and specificity, and is competitive with state-of-the-art supervised methods while being much faster to compute. The method also provides explanations for the assigned ICD codes by extracting the most informative terms from the clinical note, which aids interpretability.

Some of the key gaps and limitations were identified in the study. The method assumes a fixed vocabulary of words for computing the word embeddings. This may limit its ability to handle new, unseen medical terminologies that may appear in clinical notes. The paper focuses only on the unstructured text in clinical notes, while structured data like lab results, patient demographics, etc. could also provide useful signals for ICD code assignment.

A study by Wu et al. (2022) proposes a Joint Attention Network (JAN), an automated system to categorize International Classification of Diseases (ICD) from clinical text documents. JAN makes use of document-based attention and label-based attention to capture semantic information from clinical text and label descriptions, which helps with the classification of both frequent and rare codes, even in the long-tailed label distribution. JAN introduces an adaptive fusion layer that combines the advantages of both attention mechanisms, and CorNet blocks are used to exploit label co-occurrence relations. The results from experiments conducted on the MIMIC-III and MIMIC-II datasets show that JAN outperforms previous state-of-the-art methods. JAN achieved higher micro-F1, micro-AUC, and precision scores.

Ablation studies confirm the importance of each component. Visualizations of attention weights and label correlations help interpret the model. The key innovations of JAN are the use of label descriptions to aid in the classification of rare codes via label-based attention, the adaptive combination of document and label attention, and the capturing of label co-occurrences with CorNet blocks. The proposed JAN model effectively handles the long-tailed label distribution and label correlations, which were often ignored before.

However, there are some potential gaps or limitations that the paper identifies. Firstly, the paper does not provide a thorough analysis of the model's computational complexity and scalability, which could be important for large-scale deployment in real-world scenarios. Secondly, the error analysis section only provides some insight into the types of errors made by the model, but a more comprehensive and quantitative analysis could help better understand the model's strengths and weaknesses. Thirdly, while the paper argues that using label descriptions is more effective than leveraging label hierarchies (as done in HyperCore), it does not explore whether combining both label descriptions and hierarchies could further improve performance. Fourthly, although attention visualizations are provided, the interpretability aspect could be further strengthened by providing more concrete examples and insights into how the model arrives at its predictions, especially for challenging cases. Lastly, the paper does not compare JAN's performance with recent transformer-based models like BERT, which have shown strong results on many NLP tasks and could serve as additional baselines.

Three transformer-based model architectures were used to automate ICD coding from clinical notes in a study conducted by Liu et al. (2023). The PLM-ICD model, which was the current state-of-the-art for ICD coding, was optimized by using longer input sequence lengths and achieved significant performance improvements on the MIMIC-III and MIMIC-II datasets. However, the XR-Transformer model, which is state-of-the-art for general extreme multi-label text classification tasks, did not perform well on the ICD coding task. The paper proposes a new model called XR-LAT, which recursively fine-tunes a transformer on a hierarchical code tree using label-wise attention, knowledge transfer, and dynamic negative sampling. The XR-LAT variants achieved competitive performance compared to the optimized PLM-ICD models. The study found that handling long text sequences by segmenting them into smaller chunks performed better than using long-sequence transformers like BIGBIRD for ICD coding. On the MIMIC-III dataset, the optimized PLM-ICD model achieved a new state-of-the-art micro-F1 score of 60.8%, while an XR-LAT variant improved the macro-AUC by 2.1% over the previous best. In summary, the study optimized existing transformer models and proposed a new hierarchical recursive model to tackle the challenges of extremely large code sets and long text sequences in automated ICD coding.

There were some gaps identified in the paper. The experiments were conducted on the MIMIC-III and MIMIC-II datasets which have 8,929 and 5,031 ICD-9-CM codes respectively. This is much smaller than the full ICD-10-CM/PCS code set which has 141,747 codes. The authors note there is a need to evaluate these transformer-based approaches on datasets with a more comprehensive set of ICD codes. The segmentation method used to handle long text sequences splits the text consecutively into chunks, potentially breaking important context at the edges of each chunk. The paper suggests investigating overlapping adjacent chunks as a possible solution. ICD codes that were absent from the training datasets were not considered during model training. Integrating few-shot and zero-shot learning capabilities could help handle missing codes. While hyperbolic embeddings and asymmetric loss improved macro-F1 by emphasizing rare codes, more research is needed to further improve performance on long-tailed code distributions with many infrequent codes. The paper only explored the segmentation approach for long texts up to 5,120 tokens. Longer sequences beyond this may require additional handling.

A study by Ji et al. (2021) found that pretraining language models on corpora from domains closer to the medical/clinical domain (e.g. biomedical articles, clinical notes) improves their performance on medical code assignment compared to pretraining on general domains like books and Wikipedia. Additionally, a hierarchical fine-tuning architecture that segments long clinical notes into shorter sequences and uses an additional transformer layer helps better encode the long documents compared to simply truncating the notes.

Incorporating a label-wise attention mechanism that connects the document representation with label information further improves performance. However, despite careful architecture engineering, fine-tuning pre-trained language models still underperforms a carefully trained classical CNN model on frequent medical codes from the MIMIC-III dataset. The results suggest focusing more on improving performance for infrequent medical codes where all models struggle.

Overall, the paper demonstrates that while pre-trained language models transfer knowledge beneficially, classical models with appropriate training can still outperform transformer-based approaches like BERT on certain tasks and datasets. The study provides practical guidance for building robust medical code assignment systems.

PLM-ICD is a framework that uses pre-trained language models (PLMs) to automatically code clinical notes with ICD (International Classification of Diseases) codes and was introduced in a study by C.-W. Huang et al. (2022). The paper discusses the challenges of applying PLMs to ICD coding, such as lengthy input text, a large label set, and domain mismatch between general PLM pretraining and clinical text. To overcome these challenges, the paper proposes strategies such as domain-specific pretraining of PLMs on biomedical/clinical text, segment pooling to handle long input by splitting into segments, and label attention mechanism to deal with large label sets. The paper also highlights the state-of-the-art or competitive performance achieved on the MIMIC-III and MIMIC-II datasets using the proposed PLM-ICD framework. Additionally, the paper analyses factors that affect PLM performance, such as pretraining corpus, vocabulary, and optimization hyperparameters. The paper provides best practices for applying PLMs to ICD coding and similar tasks.

The paper identifies some gaps and trends in this area. Although the paper focuses on ICD-9 coding, it does not explore ICD-10 coding, which has a much larger label set. The paper also does not consider leveraging the hierarchical structure and code descriptions of ICD codes, which could potentially improve performance. The analysis is limited to English clinical notes, and extending PLM-ICD to other languages and multi-lingual settings could be a future direction. The paper does not explore the trend of using large language models like GPT-3 for few-shot or zero-shot learning on downstream tasks for ICD coding. Finally, more advanced architectures like Longformer, Reformers, etc., that can handle longer input sequences, could help overcome the long input text challenge more effectively.

Overall, the paper presents a solid framework for ICD coding with PLMs and provides insightful analysis, while also identifying some potential gaps and future research trends in this area.

Explainable ICD Coding

Wiegreffe et al. (2019) explores methods to leverage clinical concept extraction tools, such as the Apache cTAKES system, to improve the performance of deep learning models on the task of document-level clinical coding. The paper proposes two novel approaches to incorporate cTAKES concept annotations into the input representation for a state-of-the-art clinical coding model. The first approach is to treat the extracted concepts as features, using them to supplement or replace the raw text input. The second approach is to treat the extracted concepts as labels, using a multi-task learning setup to learn a better representation of the text. However, neither of these approaches was found to improve performance over the baseline model that uses raw text alone. The paper conducts an in-depth error analysis is conducted to understand the negative results. It finds that the cTAKES annotations do not capture a significant amount of lexical variation, as most concepts are only mapped to a single-word phrase. Additionally, the raw cTAKES code predictions were found to have limited accuracy, which may have negatively impacted the multi-task learning approach. They perform ablation studies to further investigate the contributions of the cTAKES NER and ontology mapping components. The results suggest that the positions in the text annotated by cTAKES are valuable, but the actual concept mappings do not improve performance.

In conclusion, the paper demonstrates that integrating existing clinical concept extraction tools like cTAKES does not straightforwardly improve performance on the document-level clinical coding task and provides insights into potential directions for future research.

The paper identifies a few key gaps and limitations in the approaches explored. The error analysis found that most concepts in the cTAKES annotations were only mapped to a single-word phrase. This suggests the annotations may not be effectively capturing lexical variation. The paper hypothesizes this could be a reason why the augmented representations did not improve performance over the raw text baseline. The paper found that the raw cTAKES code predictions had limited accuracy, with the top-level codes often being assigned instead of the more specific gold-standard codes. This was identified as a potential reason why the multi-task learning approach, which aimed to leverage the cTAKES predictions, did not improve performance. While the multi-task learning approach was designed to help transfer the domain knowledge encoded in cTAKES to the main clinical coding task, the paper found it did not lead to performance gains. This suggests challenges in effectively leveraging the information from the clinical concept extraction tool.

The paper by López-García et al. (2023) focuses on using transformer-based models for explainable clinical coding tasks, achieving state-of-the-art performance. Two methodologies were developed - a multi-task approach and a hierarchical-task approach. The hierarchical task approach showed superior performance.

The paper systematically analyses the performance of transformer-based models for explainable clinical coding, comparing general-domain and clinical-domain versions. It highlights the benefits of the hierarchical-task approach, demonstrating improved performance by adapting transformers to the medical domain and addressing the medical entity recognition (MER) and medical entity normalization (MEN) tasks separately.

The research compares the multi-task and hierarchical-task approaches for MER and MEN, showcasing the effectiveness of the hierarchical-task strategy in reducing complexity. Overall, the paper focuses on improving automatic clinical coding by making it more explainable, with the hierarchical-task approach outperforming the multi-task strategy and setting new performance standards that could be applicable to other clinical tasks involving medical entity recognition and normalization.

Cheng et al. (2023) recently introduced MDACE, which is the first publicly available dataset that contains evidence annotations for medical codes from clinical notes. It was built on a subset of the MIMIC-III clinical records dataset and includes 302 inpatient charts and 52 outpatient charts that were annotated by professional medical coders. These coders provided evidence text spans that support assigned medical diagnosis and procedure codes. The annotations cover both ICD-9 and ICD-10 codes, with ICD-10 codes mapped back to ICD-9 to enable use with both MIMIC-III and the newer MIMIC-IV datasets.

Creating MDACE involved overcoming several challenges, such as different coding specialities (inpatient vs outpatient), mapping between code systems, and ensuring consistent evidence annotation. The paper establishes baseline performance on evidence extraction from the MDACE dataset using variants of a convolutional attention model. MDACE enables automatic evaluation of evidence/rationale extraction for medical coding systems, as well as interpretability evaluation for multi-label text classification models on long documents. In summary, MDACE is a new annotated dataset released to promote research on extracting supporting evidence for predicted medical codes from clinical notes, which is an important step towards more interpretable medical coding systems.

While discussing the limitations of the MDACE dataset, the authors note that professional coders are trained to find sufficient evidence for each code, rather than providing exhaustive evidence. For larger patient encounters, Profee coders may have missed some evidence due to time constraints. The recall metric on this dataset may underestimate the true recall of evidence extraction systems since inpatient coders annotate only sufficient evidence. During data cleanup, the authors observed some inconsistencies and human errors in the annotations. Coders sometimes annotated only partial evidence, leaving out modifiers like "acute", "moderate", "bilateral", etc. The annotation tool also did not allow highlighting and linking discontinuous text spans as single evidence, leading to some evidence spans containing extra tokens or missing parts of the evidence. Although the authors attempted to fix these issues, some errors may still exist in the dataset's annotations.

As an expert-annotated dataset, MDACE has the inherent limitation of being relatively small in size compared to other datasets. In summary, the main limitations relate to potential incompleteness and noise in the evidence annotations due to the complexity of the annotation task, coder disagreements, and the constraints of the annotation process and tools used.

The literature review provides a comprehensive overview of the existing research on automating medical coding using natural language processing (NLP) techniques. Several key themes and connections emerge across the different studies:

1. Evolution from rule-based to machine learning approaches: Early work relied on handcrafted rules and lexicons, but more recent studies have shifted towards data-driven machine learning models, particularly deep learning architectures like neural networks and transformers. This transition was enabled by the availability of large annotated clinical datasets and advancements in deep learning.

2. Leveraging pre-trained language models: Multiple studies Patel et al. (2017), Ji et al. (2021), Huang et al. (2022) and Liu et al. (2023) explored the use of pre-trained language

models like Word2Vec, BERT, and their domain-specific variants. These models capture general linguistic knowledge and can be fine-tuned on clinical data, outperforming models trained from scratch.

3. Handling challenges in medical coding: Several challenges were addressed, including long input text sequences Liu et al. (2023) and Ji et al. (2021), large label spaces Huang et al. (2022) and Wu et al. (2022), infrequent/rare codes Xie & Xing (2018) and Ji et al. (2021), and long-tailed label distributions Wu et al. (2022). Proposed solutions involved techniques like segmentation, label-wise attention, hierarchical modelling, and adversarial training.

4. Incorporating domain knowledge: Studies explored different ways to incorporate domain knowledge, such as adapting pre-trained embeddings with medical coding information Patel et al. (2017), utilizing code hierarchies Xie & Xing (2018), and leveraging label descriptions Wu et al. (2022).

5. Classical models vs. transformers: While transformer-based models showed strong performance, some studies Ji et al. (2021) found that classical models like CNNs could still outperform transformers on certain datasets and code frequencies, highlighting the importance of appropriate model selection and training strategies.