MULTI-LABEL CLASSIFICATION OF MEDICAL TEXT USING CLASSIFIER CHAINS

Maria Marouli

Department of Computer Science and Biomedical Informatics
University of Thessaly

Supervisor: Associate Prof. Sotirios Tasoulis

Lamia, June 2025

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Declaration

I declare that this thesis is my own work and has not been submitted elsewhere.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Abstract

Healthcare systems rely on the determination of International Classification of Diseases (ICD) codes for their operation. So far, medical code tagging required a specialist to read large amounts of unstructured medical text, making it a time-consuming and labor-intensive approach. Recently, many approaches have been proposed that involve using state-of-the-art Natural Language Processing (NLP) techniques, to automate this process and reduce costs. Inspired by the success of Deep Learning and Classifier Chains - a mathematical model that takes into account label inter-dependencies – we present a novel approach that integrates these two methods in order to automate the multi-label classification of medical text. Our approach explicitly models label inter-dependencies, a key characteristic of ICD codes, and demonstrates its impact on improving classification performance. Using discharge summaries from the MIMIC-III database, we implement an application-driven pipeline combining advanced natural language processing (NLP) techniques with ensemble methods to optimize chain order. Comparisons with both traditional and advanced multi-label classification models demonstrate the practical benefits of our approach in terms of accuracy and computational efficiency. These findings validate the initial hypothesis and effectively connect machine learning research with practical, real-world applications.

Keywords: ICD-9, Neural Networks, Classifier Chains, Multi-Label Classification, MIMIC-III

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Contents

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 1

Introduction

Electronic Health Records (EHRs) serve as essential tools within the healthcare system to document the history of the patient and record important clinical information. EHRs contain a multitude of structured and unstructured data, from discharge summaries to biochemical results and imaging technologies. These data should be coded according to the International Classification of Diseases (ICD) nomenclature created by the World Health Organization (WHO) to facilitate hospital management and billing purposes, as well as to create a valid record of patient history. The ICD-9 codes consist of 5 digits: the first three digits represent a generalized medical concept, and the last two digits represent specific variations of that concept, thus forming a hierarchy.

The rapid digitalization of medical content highlights the growing demand for automated ICD coding systems to streamline the organization and utilization of medical information. Assigning ICD codes presents notable challenges due to the complex relationships among codes, their hierarchical structure, and the unstructured nature of the clinical text, making it a vital area for research and innovation [30]. The research community has proposed several methods for tackling this complex yet crucial task - including deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) [25], transformer-based architectures with label attention mechanism [7], as well as deep transfer learning [32] - that are able to exhibit exceptional results given the nature of this problem.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

In this work, we utilize Classifier Chains along with an Artificial Neural Network as base classifier [26], on the task of predicting the ICD-9 coding. To achieve optimal results using Classifier Chains, we implemented an analysis for the optimal chain order to exploit the inter-dependencies found in the ICD-9 coding. The effectiveness of this approach is then compared with well-established and state-of-the-art methods.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 2

Literature Review

For more than two decades, automatic ICD-9 coding has been an ever-growing area of research. This task can be seen as a multi-label classification problem, where each ICD-9 code is a label, and each patient has multiple ICD-9 codes. There are two major approaches for automatically assigning ICD-9 codes to free-text medical content. The rule-based approach to ICD-9 coding involves the development of systems that rely on predefined if-then rules derived from coding guides and expert knowledge [5]. Their advantages are the independence of labeled training data and the direct incorporation of medical guidelines and domain expertise. However, the time-consuming nature of manually defining rules, challenges handling large code sets, and difficulties adapting to rare synonyms or institution-specific language variations pose significant limitations on their use.

On the other hand, learning-based systems do not rely on external medical knowledge, offering higher scalability and maintainability. Models used in this approach are based on well-known machine learning methods that are able to find the underlying distribution of the provided datasets, as in [6, 17, 33].

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Deep Learning has also been extensively applied to this multi-label classification problem. Proposed methods include utilizing state-of-the-art neural network architectures to solve this complex task. Such as, [9] introduced a Convolutional Neural Network (CNN) architecture to automate this task, [18] exploited raw text from Wikipedia as a source of knowledge and presented condensed memory neural networks to learn the diagnosis on MIMIC-III data. [1] introduced a two-stream deep learning model for ICD-9 prediction, leveraging both numerical and text-based data. Furthermore, [16] presented a Bidirectional LSTM (bi-LSTM) model with an attention layer to assign medical text to ICD-9 codes.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 3

Multi-Label Classification

Multi-label classification, a variant of the traditional classification problem, is a widely used predictive task in real-world problems such as document tagging and image annotation. In multi-label problems classes are not mutually exclusive, allowing each instance to belong to multiple classes at the same time.

A multi-label dataset can be denoted $D = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{N}$, consisting of N examples. Each i-th instance $\mathbf{x}^{(i)}$ is associated with a label vector $\mathbf{y}^{(i)} = [y_1^{(i)}, \ldots, y_L^{(i)}]$; there are L elements corresponding to L label concepts, and each element $y_j^{(i)} \in \{0, 1\}$ indicates the relevance (if 1) or not (if 0) of the j-th concept to this instance. A multi-label model is tasked with providing predictions $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_L]$ for any given test instance $\tilde{\mathbf{x}}$. Note that traditional multi-class learning also considers L concepts but, in contrast, only a single concept can be relevant to any particular instance.

A common approach to multi-label classification is to perform *problem transformation* [29], which involves decomposing the task into one or more single-label (i.e. binary, multi-class) problems. In this way, the final multi-label prediction is determined by aggregating the results of all the single-label classifiers. Problem transformation is attractive due to its scalability and flexibility, since any single-label classifier can be used.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

The most common problem transformation method is called the *binary relevance* method. Binary Relevance (BR) breaks the initial problem into multiple binary problems, where each binary classifier learns to predict the relevance of one of the labels. BR is theoritically simple and intuitive and scales linearly with the number of labels. However, it does not take into account label correlations, limiting its predictive performance.

A second problem transformation method is the binary *pairwise* classification approach (PW), where a binary classifier is trained for each pair of labels. However, PW faces quadratic complexity in terms of the number of labels, making intractable for large problems.

Label Power-Set (LC) is another approach that can be applied to multi-label classification as a problem transformation method. LC transforms a multi-label problem to a multi-class problem by treating each label set as a single label. So, LC assigns a unique class to every possible label combination that is present in the training data. While LC is able to capture label correlations, it has a worst-case computational complexity (exponential with the number of labels) and tends to over-fit the training data.

The final problem transformation approach is the *Classifier Chains* method [21], which extends the BR method. Classifier Chains does not only train a binary classifier for each label but is also able to pass label information between classifiers through the feature space. It thus combines the computational efficiency of the BR method, while also accounting for label inter-dependencies which are critical for classification.

The alternative to problem transformation is called *algorithm adaptation* [29], where an algorithm is directly modified in order to make multi-label predictions. Popular algorithms used in this approach include AdaBoost and Decision Trees.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 4

Methodology

In this study, various machine learning and deep learning models are evaluated, although the main scope is to compare the Classifier Chains method against traditional multi-label text classification methods, in order to validate the initial hypothesis that label inter-dependencies do play a significant role in medical text classification tasks. More precisely, the focus is on classifying unstructured discharge summaries from the MIMIC-III database. Due to the complexity of the prediction task, given the common rarity of ailments and rare ICD-9 codes, mitigating the number of ICD-9 codes to the most frequent was deemed necessary, as noted in [10]. Therefore, the number of ICD-9 codes used in this study limited to twenty. This choice aims to improve performance and training time while reducing computational costs, allowing for a more efficient evaluation of the proposed method.

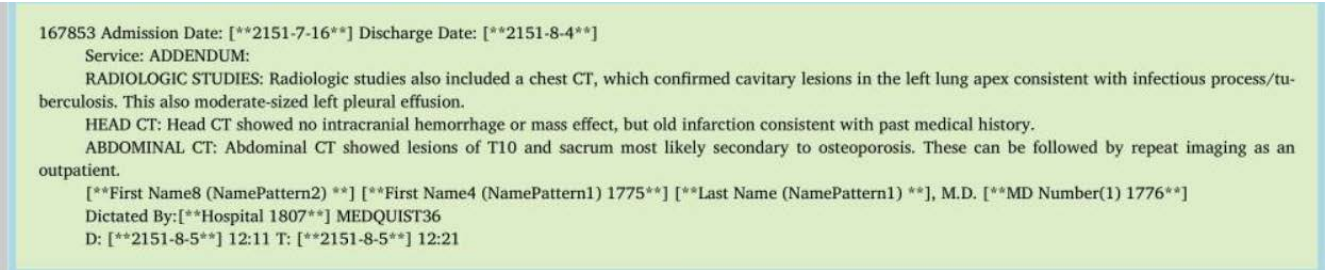
Additionally, our study aims to compare our results to state-of-the-art methodologies that apply to this particular multi-label classification problem. Specifically, we follow the methodology proposed in [10], where different machine learning (Logistic Regression, Random Forest) and deep learning (Recurrent Neural Networks, Long Short-Term Memory Networks, Gated Recurrent Unit Networks) were used to classify unstructured discharge summaries to the 10 most frequent ICD-9 codes, using the MIMIC-III data. Our study assesses whether leveraging Classifier Chains can lead to improved performance, thus providing a thorough comparative analysis.

The following sub-sections will provide a detailed description of the methodology followed in this study.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.1 Data

MIMIC III [11] is a publicly available database that consists of de-identified records from Birth Israel Deaconess Medical Center intensive care unit (ICU) stays, collected between 2001 and 2012. The database includes 52,726 free-text discharge summaries and covers 942 distinct ICD-9 categories. The objective of this study is to use unstructured medical text and implement various Natural Language Processing (NLP) techniques, to address the problem of multi-label classification while taking into account label inter-dependencies. For this purpose only the *noteevents* table was used and the focus was on the *discharge summaries* category. A short example of these discharge summaries can be found in Figure 1 The discharge summaries are typically much longer.



167853 Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**]
Service: ADDENDUM:
RADIOLOGIC STUDIES: Radiologic studies also included a chest CT, which confirmed cavitary lesions in the left lung apex consistent with infectious process/tuberculosis. This also moderate-sized left pleural effusion.
HEAD CT: Head CT showed no intracranial hemorrhage or mass effect, but old infarction consistent with past medical history.
ABDOMINAL CT: Abdominal CT showed lesions of T10 and sacrum most likely secondary to osteoporosis. These can be followed by repeat imaging as an outpatient.
[**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**]
Dictated By:[**Hospital 1807**] MEDQUIST36
D: [**2151-8-5**] 12:11 T: [**2151-8-5**] 12:21

Figure 1: Sample Discharge Summary from MIMIC-III Dataset

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.2 Data Preprocessing

The data preprocessing pipeline contains various steps that are crucial for transforming the dataset in a way that prepares it for classification, while ensuring alignment between the clinical text and the diagnostic labels.

First of all, the ICD-9 codes were grouped into their major 3-digit categories. This was done by removing the last one or two digits and keeping only the first three. This grouping helped to focus on broader diagnostic categories.

Secondly, the 20 most frequent ICD-9 categories were kept, in order to minimize the computational complexity of this multi-label classification problem.

It was observed that for each unique hospital admission there were more than one discharge summaries recorded and stored in the database. To avoid duplication of medical text, only the first discharge summary was retained.

In the end, the final matrix was created, where the independent variable consists of the unstructured discharge summaries and the dependent variable represents the ICD-9 categories. Each discharge summary is associated with one or more ICD-9 categories, forming a binary matrix.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.3 Text Preprocesing

Several preprocessing techniques were applied in order to clean and transform the unstructured text data in a way that is suitable for downstream NLP tasks such as text classification [3].

Firstly, all words were converted to lowercase in order to eliminate case sensitivity. Special characters, numbers, brackets and their content, as well as punctuation were discarded since they do not affect the core meaning of the text content. Subsequently, placeholders corresponding to personal information were also excluded from the text data.

Finally, stop words like “and” and “a” were filtered out since they hold little significance for traditional machine learning models. However, when building deep learning models these common words were retained, since they may be significant to the contextual understanding of the text. A typical text preprocessing pipeline is depicted in Figure 2.

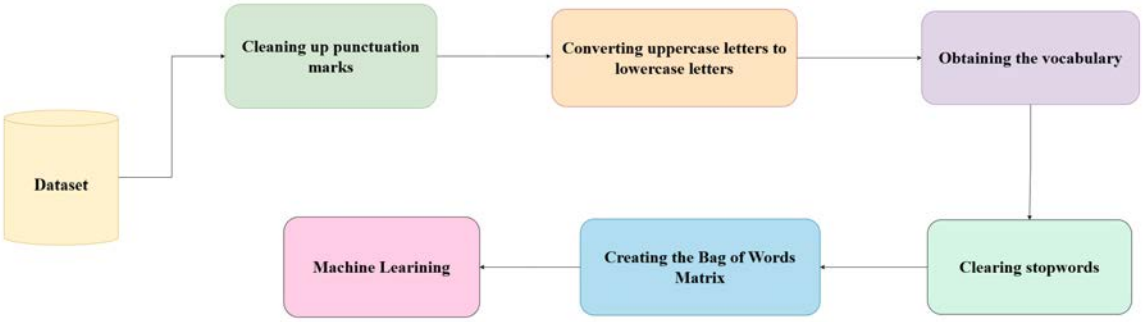


Figure 2: Text Preprocesing

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.4 Feature Extraction

For feature extraction, two approaches were employed: Term Frequency – Inverse Document Frequency (tf-idf) [23] and BioWordVec [31].

The tfidf is a statistic that reflects the significance of a word within a document, relative to a collection of documents (corpus). Tfidf is calculated as the product statistical terms: TF (Term Frequency) and IDF (Inverse Document Frequency). TF measures the frequency of a term within a document, while IDF quantifies the rarity of a word across the corpus. As a result, common words in the corpus are assigned lower weights, while rarer words receive higher weights. The document-term matrix was created utilizing the 20,000 most frequent words in the entire corpus.

The second approach to feature extraction is utilizing the pretrained word embeddings generated by the BioWordVec model. In order to build the embedding matrix that will be fed to deep learning models, all text samples in the training data were first tokenized. Then the 10,000 most frequent words in the corpus were mapped to their embedding vectors provided by the BioWordVec model.

It is important to note that only the LSTM model uses the BioWordVec embeddings, while the other models rely on tf-idf to convert text data into numerical representations.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.5 Models

The following sub-sections will provide a detailed description of the models and their implementation in our study.

4.5.1 Baseline Models

Logistic Regression [27] is a supervised machine learning technique, widely adopted to the task of classification. It is used to estimate the relationship between a dependent categorical variable and one or more independent ones. LR models seek to minimize the negative log likelihood function (4.1) using the process of gradient descent to find global maximum.

$$l(\theta) = -\sum_{i=1}^{n} \left(y_i \log \hat{y}_{\theta,i} + (1 - y_i) \log(1 - \hat{y}_{\theta,i})\right) \tag{4.1}$$

Dealing with a multi-label classification task we employed the technique of Binary Relevance (BR). Under BR, a one-vs-all strategy, $|L|$ separate classifiers were trained, one for each label.

Long Short-Term Memory (LSTM) [8] as shown in Figure 3, is a type of Recurrent Neural Network (RNN), used in the field of deep learning for text classification tasks. LSTMs contain memory cells that allow them to retain information over long sequences. Their ability to selectively remember or forget information, enables them to mitigate problems of vanishing and exploding gradients, leading to more effective training and better capturing of long term patters in sequential data. The LSTM architecture used in this study consists of an Embedding Layer followed by a LSTM Layer with 32 hidden units, a Dropout Layer, a LSTM Layer with 16 hidden units and a Dense Layer with 20 output units. The model uses a sigmoid activation function for the output layer, which is suitable for multi-label

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

classification tasks, and was trained with the Adam optimizer and the binary cross-entropy loss function (4.2).

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log(p(y_i)) + (1-y_i)\cdot \log(1-p(y_i))\right) \tag{4.2}$$

To tackle the task of classifying a multi-label problem with the LSTM, algorithm adaptation method was employed. Since neural networks are inherently flexible, they can be adapted for multi-label classification by modifying the output layer. Specifically, each neuron in the output layer corresponds to a different label, allowing the network to predict multiple labels at once [29].

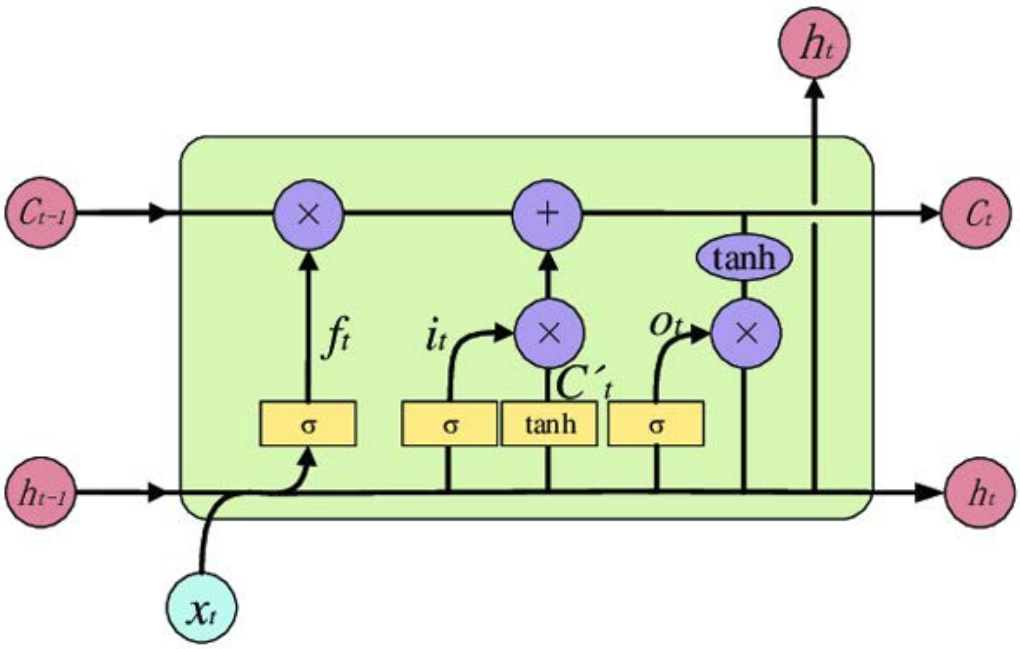


Figure 3: LSTM Neural Network Architecture
Source: https://www.proxpc.com/blogs/5-types-of-lstm-recurrent-neural-networks

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.5.2 Classifier Chains

Classifier Chains (CC) [20] is a machine learning method for problem transformation in multi-label classification. It is closely related to Binary Relevance (BR), since both methods train a single binary classifier per label. However, Classifier Chains extends Binary Relevance by including label inter-dependencies during training and inference.

The binary classifiers in the CC model are linked along a chain, where each classifier deals with a single binary relevance problem associated with label $l_j \in L$. Hence a chain $C_1, \ldots, C_{|L|}$ of binary classifiers is formed. Each classifier C_j in the chain is responsible for learning and predicting the binary association of label l_j given the feature space, augmented by all prior binary relevance predictions in the chain $l_1, \ldots, l_{j-1}$. The classification process begins at C_1 and propagates along the chain: C_1 determines the conditional probability $P(l_1 \mid \mathbf{x})$ and every following classifier $C_2, \ldots, C_{|L|}$ predicts $P(l_j \mid \mathbf{x}_i, l_1, \ldots, l_{j-1})$.

This chaining passes label information through the classifiers, allowing CC to take into account label information. This means that each binary classifier uses the preceding labels as additional explanatory variables, overcoming the label independence problem of the BR method. This effectively allows the Classifier Chains method to model relationships between labels.

$\boldsymbol{X}$	Y_1
$\boldsymbol{x}^{(1)}$	0
$\boldsymbol{x}^{(2)}$	1
$\boldsymbol{x}^{(3)}$	0
$\boldsymbol{x}^{(4)}$	1
$\boldsymbol{x}^{(5)}$	0
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$

$\boldsymbol{X}$	Y_1	Y_2
$\boldsymbol{x}^{(1)}$	0	1
$\boldsymbol{x}^{(2)}$	1	0
$\boldsymbol{x}^{(3)}$	0	1
$\boldsymbol{x}^{(4)}$	1	0
$\boldsymbol{x}^{(5)}$	0	0
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$	$\hat{y}_2$

$\boldsymbol{X}$	Y_1	Y_2	Y_3
$\boldsymbol{x}^{(1)}$	0	1	1
$\boldsymbol{x}^{(2)}$	1	0	0
$\boldsymbol{x}^{(3)}$	0	1	0
$\boldsymbol{x}^{(4)}$	1	0	0
$\boldsymbol{x}^{(5)}$	0	0	0
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$

$\boldsymbol{X}$	Y_1	Y_3	Y_3	Y_4
$\boldsymbol{x}^{(1)}$	0	1	1	0
$\boldsymbol{x}^{(2)}$	1	0	0	0
$\boldsymbol{x}^{(3)}$	0	1	0	0
$\boldsymbol{x}^{(4)}$	1	0	0	1
$\boldsymbol{x}^{(5)}$	0	0	0	1
$\tilde{\boldsymbol{x}}$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$

Figure 4: The transformation of a dataset for the application of classifier chains. It is worth highlighting that the 0/1 values shown in the tables correspond to the values $y_j^{(i)}$ found in the training data; whereas $\hat{y}_1, \ldots, \hat{y}_L$ (in the final row) are all predicted values not available at training time [22].

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

4.5.3 Classifier Chains as Neural Networks

Another way to model the Classifier Chains method is along with a neural network architecture, as proposed by [26]. In this approach, Artificial Neural Networks (ANNs), Figure 5, are used as base classifiers, where each classifier is trained independently using the input features $\mathbf{x}$ and ground-truth values, similarly to the traditional CC model. The model forms a chain, where each classifier uses the preceding predictions as additional features to predict whether an instance belongs to a particular label. The output of an ANN is fed as input to the next, hence maintaining the interdependency consideration.

While this is a promising approach, there are major drawbacks. Firstly, each ANN is trained independently, thus back-propagation occurs individually. This can lead to inefficient dependency modeling, since the label inter-dependencies are only captured implicitly through the structure of the chain, similar to the classic CC method. Secondly, there is an issue regarding the direction in which the dependency is considered in the chain architecture. Each label l_j is considered dependent on its preceding label, but not its succeeding.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Although, this drawback can be mitigated by training two models from both directions and aggregating their results to get better predictions, as proposed by [26], in this work we only make use of their proposed uni-directional approach for training the Classifier Chains Neural Network architecture due to its simplicity and straightforward evaluation. In detail, this architecture consists of four Dense Layers interconnected with Dropout layers to mitigate over-fitting.

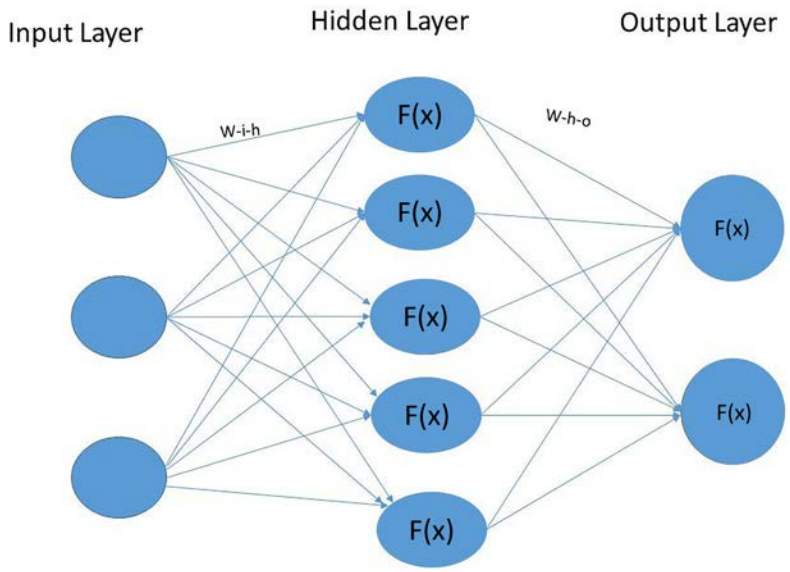


Figure 5: Artificial Neural Network Architecture
Source: https://humboldt-wi.github.io/blog/research/

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 5

Optimal Chain Order

The performance of Classifier Chains [24] is heavily affected by the order of labels in the chain. Since, at prediction time, each classifier in the chain relies on the predictions of its preceding classifiers, a poor chain order can activate the problem of error propagation: incorrect predictions made early in the chain can subsequently be propagated or even reinforced along the entire chain. However, finding an optimal chain order is a problem of combinatorial complexity, due to the factorial growth of possible permutations. This fact drives researchers to rely on heuristics or approximate search methods to make the problem tractable [22].

A recently proposed method is called the Ensemble Classifier Chains (ECC) [20]. ECC generates multiple chains of random order and then combines the predictions through averaging/majority voting, reducing the error variance caused by the randomness. However, this strategy mitigates poor chain orders rather than finding an optimal order. Another approach is to apply a heuristic to label dependence [13, 12]. This approach measures label interdependence, a highly studied problem for which several tools exist, and then uses these measurements to create an order. A third approach involves designing heuristics around base classifier accuracy [13, 4]. In this method, labels that are easier to predict are placed near the beginning of the chain since it is supposed that they are less likely to generate errors that will then be propagated down the chain.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Due to the high computational disadvantages of evaluating different label orders and the question of whether there is a global optimum chain order, a different method to find the optimal chain structure is proposed by [15]. They introduced a novel method called the Dynamic Classifier Chains (DCC), based on Random Decision Trees (RDTs), which can dynamically choose the label ordering for each unique prediction. Because RDTs do not optimize an objective function during training, [15] adapted the Extreme Gradient Boosted Trees (XGBOOST) algorithm to the DCC setting (XDCC). The main difference between RTDs and XDCC and the traditional CC model is that the next label chosen to be predicted for each instance can be different.

In our work, we implement an Ensemble Classifier Chains model [20] in order to achieve better classification results by leveraging different permutations of label orders. Additionally, we treat this model as a heuristic method to determine an optimal label order based on accuracy. We retrain both the CC and ANN-CC models using this label order to further enhance their predictive power.

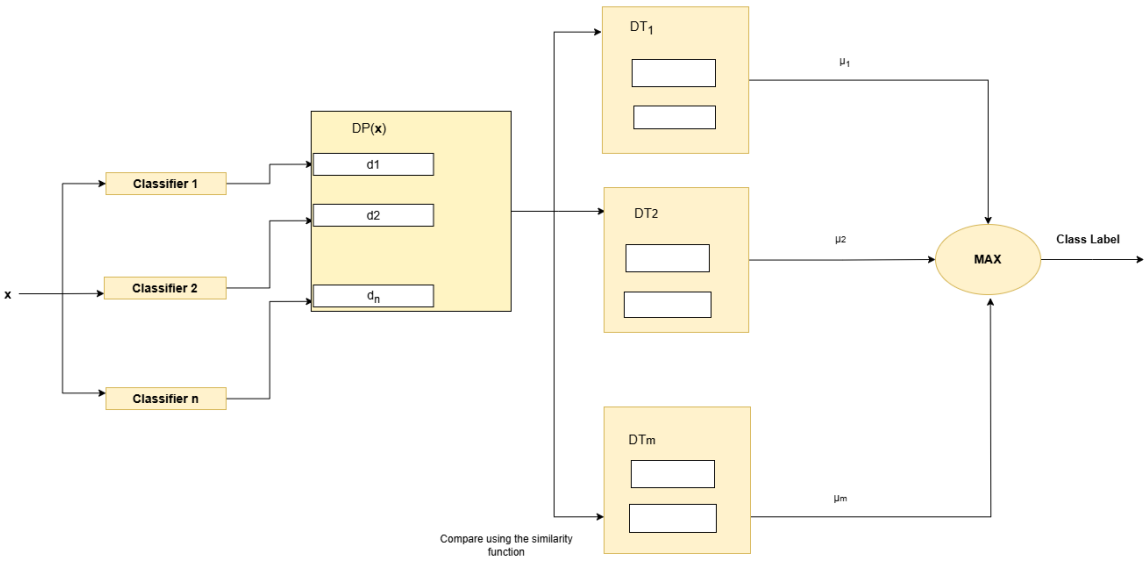


Figure 6: Ensemble Classifier Chains

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 6

Evaluation Metrics

Several metrics were utilized to comprehensively evaluate the performance of models in the multi-label classification task. These metrics, as defined in [2], include Accuracy, Jaccard Score, F1-Score, Precision, and Recall. Each metric provides unique insights into model performance.

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i \cap Y_i|}{|Z_i \cup Y_i|} \tag{3}$$

$$\text{Jaccard Score} = \frac{|Z_i \cap Y_i|}{|Z_i \cup Y_i|} \tag{4}$$

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i \cap Y_i|}{|Z_i|} \tag{5}$$

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z_i \cap Y_i|}{|Y_i|} \tag{6}$$

$$\text{F1 Score} = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Z_i \cap Y_i|}{|Y_i| + |Z_i|} \tag{7}$$

Where Z_i is the set of true labels, Y_i is the set of predicted labels, n is the number of samples and $|L|$ is the total number of labels in the label space.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Accuracy represents the overall proportion of correctly predicted labels relative to the total number of labels across all instances, offering a general measure of correctness. Precision indicates the proportion of predicted labels that are actually correct, thus assessing the accuracy of positive predictions. Recall reflects the proportion of actual labels correctly predicted, measuring the model's capability in identifying positive labels.

The F1-Score is the harmonic mean of Precision and Recall, providing a balanced metric that considers both precision and recall. It is especially valuable when dealing with imbalanced class distributions. Jaccard Score measures the overlap between predicted and true labels, highlighting the similarity between model predictions and actual labels.

By employing these metrics, we aim to evaluate the models from multiple perspectives, ensuring both overall performance and label-specific accuracy are thoroughly assessed.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 7

Summary of models

This section offers a brief overview of the models compared in this study, providing the reader with context to better understand the results.

The models used in this study include Logistic Regression (LR), Long Short-Term Memory (LSTM), Ensemble Classifier Chains Classification (ECC), Classifier Chains (CC), Artificial Neural Network-Classifier Chains (ANN-CC), Classifier Chains with Optimal Label Order (CC*), and Artificial Neural Network-Classifier Chains with Optimal Label Order (ANN-CC*).

The notation with an asterisk (*) denotes models trained with an optimal label order, which was determined by treating ECC as a heuristic method and trained using 60 different label permutations. The label order that produced the best results in terms of accuracy was selected as the optimal one.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 8

Results

In the experiments we conducted we used both the ten and the twenty most frequent ICD-9 categories as they were found in the MIMIC-III dataset. This comprehensive experimental procedure helped us in determining both the significance of inter-dependencies and the usefulness of the optimal chain order search. The results presented are the average of a statistical evaluation performed over multiple runs.

8.1 10 most frequent ICD-9 codes

In our baseline experiments, the LR model outperformed the LSTM model, achieving an *Accuracy* of 0.5392, a *Macro F1-Score* of 0.6939 and a *Macro Jaccard Score* of 0.5406, compared to 0.4797, 0.6217, and 0.4735 achieved respectively by the LSTM. This indicates that, without explicitly modeling label inter-dependencies, a simpler linear model - the LR - performed better than an LSTM network.

When label inter-dependencies were taken into account via the CC methodology, the basic CC model achieved an *Accuracy* of 0.5319 and a *Macro F1-Score* of 0.6898. In contrast, the ANN-CC model demonstrated better performance, attaining an *Accuracy* of 0.5551 and a *Macro F1-Score* of 0.7013. The Ensemble Classifier Chains model likewise performed comparably to CC — notably close, though slightly higher in *Accuracy* (0.5337) while yielding a *Macro F1-Score* of 0.6890. From this fact we can derive to the conclusion that there is great inter-dependency between

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

the ICD-9 codes even in the case where a random chain was used in the execution of the CC method.

To investigate the impact of the inter-dependence chain order, we used the ECC model as a heuristic to identify an optimal chain order based on *Accuracy*. Applying this optimal order, indicated by "$*$" in Table 8.1, led to mild performance gains for both CC ANN-CC. Specifically, CC* reached an *Accuracy* of 0.5386 and a *Macro F1-Score* of 0.6915, while ANN-CC* achieved an *Accuracy* of 0.5560 and a *Macro F1-Score* of 0.7036.

In summary, while LR outperforms LSTM when label dependencies are not explicitly modeled, incorporating these dependencies via CC generally improves results, and the deep-learning-based ANN-CC approach achieves the best overall performance. The ECC model—though helpful as a heuristic for finding an optimal chain order—exhibits performance similar with CC under default conditions. Moreover, similar observations can be seen in Figure 7, which highlights how each model compares on *Accuracy* and *Macro F1-Score*.

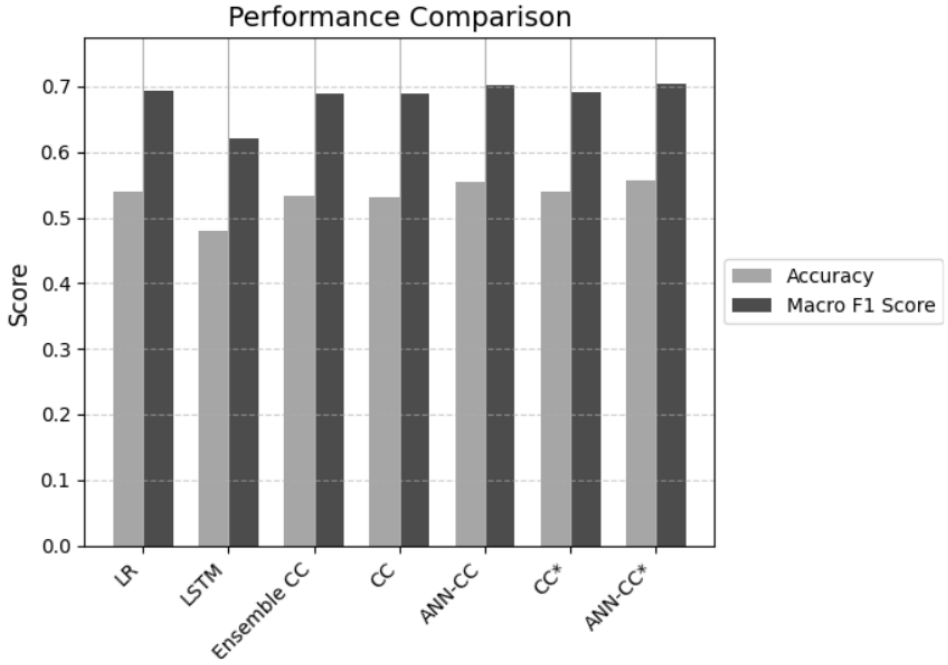


Figure 7: Performance Comaprison

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Table 8.1: Performance Comparison (10 most frequent ICD-9 Categories). Acc = Accuracy

Model	Micro JS	Macro JS	Micro F1	Macro F1	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Acc
LR	0.5426	0.5406	0.7035	0.6939	**0.8122**	**0.8102**	0.6205	0.6117	0.5392
LSTM	0.4735	0.4708	0.6425	0.6217	0.7653	0.7539	0.5542	0.5460	0.4797
CC	0.5357	0.5356	0.6988	0.6898	0.8117	0.8098	0.6135	0.6055	0.5319
ANN-CC	0.5453	0.5484	0.7058	0.7013	0.7328	0.7313	0.6807	0.6807	0.5551
ECC	0.5370	0.5351	0.6987	0.6890	0.8074	0.8043	0.6159	0.6072	0.5337
CC*	0.5408	0.5379	0.7002	0.6915	0.8039	0.8048	0.6229	0.6124	0.5386
ANN-CC*	**0.5486**	**0.5515**	**0.7085**	**0.7036**	0.7254	0.7244	**0.6925**	**0.6869**	**0.5560**

8.2 20 most frequent ICD-9 codes

The first observation in this case is that in pairs the results of the LR - LSTM, and CC - ANN-CC is the similar to the case of classifying the ten most frequent ICD-9 codes. even the studding of the ECC method is the same.

The key difference in the results of the two approaches is the behavior of the CC and ANN-CC algorithms based on the investigation or not for an optimal chain order. Specifically, While CC* slightly improves upon CC, ANN-CC* does not fully benefit from the rearranged chain order.

In terms of overall model comparisons, LR consistently outperforms LSTM. Among the methods that explicitly model label dependencies, ANN-CC once again achieves the highest *Accuracy* 0.5152 and *Macro F1-Score* 0.6422 when using the default chain order. Meanwhile, ECC remains close to CC, showing only modest gains across the majority of the evaluation metrics.

Hence, while the general hierarchy of performance is similar to that observed during classifying the 10 most frequent ICD-9 Categories, the results here highlight that an optimal chain order may not always yield consistent benefits across different neural network–based Classifier Chains. This further underscores that improvements from chain-order heuristics can be dataset-dependent and potentially architecture-dependent as well.

Similar observations can be seen in Figure 8, which highlights how each model compares on *Accuracy* and *Macro F1-Score*.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Table 8.2: Performance Comparison (20 most frequent ICD-9 Categories). Acc = Accuracy

Model	Micro JS	Macro JS	Micro F1	Macro F1	Micro Precision	Macro Precision	Micro Recall	Macro Recall	Acc
LR	0.4736	0.4528	0.6427	0.5999	**0.8065**	**0.7921**	0.5343	0.5040	0.4826
LSTM	0.3899	0.3533	0.5610	0.4633	0.7472	0.6101	0.4493	0.4104	0.4054
CC	0.4649	0.4406	0.6347	0.5867	0.8004	0.7917	0.5259	0.4914	0.4751
ANN-CC	**0.4962**	0.4911	**0.6633**	0.6422	0.7052	0.6906	**0.6261**	**0.6261**	**0.5152**
ECC	0.4699	0.4481	0.6393	0.5936	0.8060	0.7898	0.5298	0.4997	0.4782
CC*	0.4698	0.4467	0.6393	0.5927	0.8001	0.7913	0.5322	0.4988	0.4785
ANN-CC*	0.4955	**0.4916**	0.6627	**0.6430**	0.7057	0.6908	0.6246	0.6078	0.5122

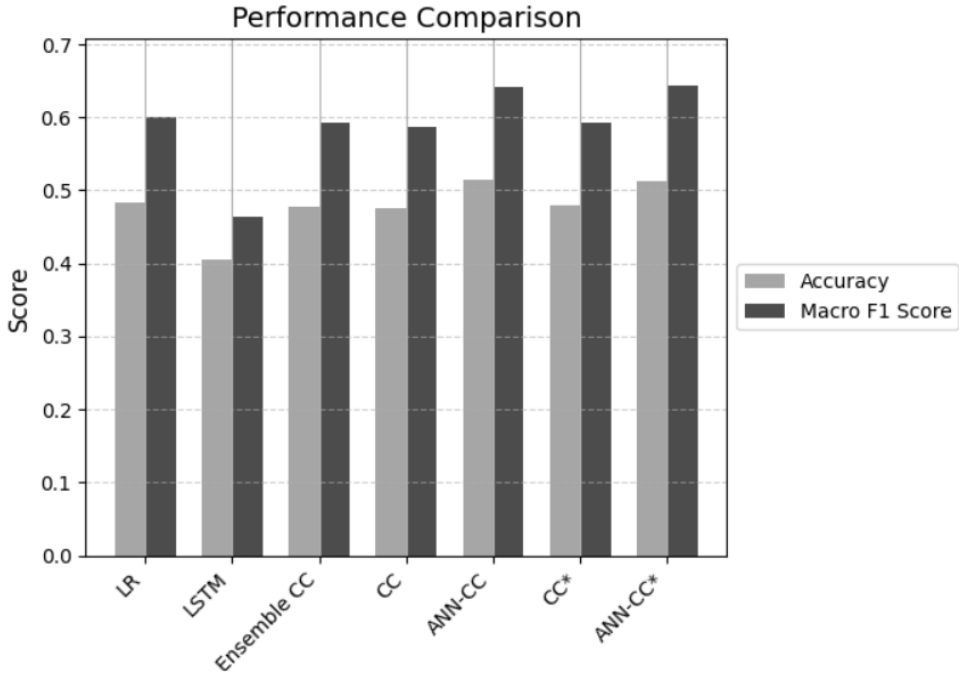


Figure 8: Performance Comaprison

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

8.3 Results Comparison

Additionally, our experiments were compared with the models proposed by [10]. TABLE 8.1 presents the results achieved by our models across the 10 most frequent ICD-9 categories, similar to [10]. TABLE 8.3 shows a comparison of our models with theirs across the following evaluation metrics: *Macro F1-Score, Macro Precision, Macro Recall* and *Accuracy*. Note that CC* and ANN-CC* indicate models trained using an optimal label order, which was determined by leveraging the ECC model as a heuristic method.

By analyzing the results presented in TABLE 8.1, we observe that the ANN-CC* model achieves better performance. This indicates that label inter-dependencies do play a significant role in solving this multi-label classification problem.

When it comes to comparing our results to those in [10], as shown in TABLE 8.3, we examine whether Classifier Chains can outperform their results. Our baseline Logistic Regression model achieves a higher F1-Score than their corresponding model, although their LSTM model still yields better results. Notably, even CC model surpasses their baseline LR model. Our best-performing model (ANN-CC*) achieves a *Macro F1-Score* of 0.7036, which is slightly lower than the score achieved by their best model (LSTM). However, ANN-CC* achieves a higher recall, indicating that modeling label inter-dependencies leads to more labels being correctly identified. Furthermore, traditional CC models achieve higher precision scores than the LSTM model proposed in [10], suggesting that incorporating label inter-dependencies may help reduce false positive predictions.

This comparison underscores that label inter-dependencies impact classifier performance in this multi-label classification problem involving the 10 most frequent labels. When leveraged, they can yield similar or slightly better results.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

In table 8.4 we compare some technical aspects between the LSTM model proposed by [10] and our best performing model, ANN-CC. The results indicate that while our model achieves a comparable Macro F1-Score, it requires fewer computational resources, making it suitable for large-scale problems.

Table 8.3: Results Comparison

Model	**Macro F1**	**Macro Precision**	**Macro Recall**	**Accuracy**
LR [10]	0.6141	0.6458	0.5856	0.7994
LSTM [10]	**0.7090**	0.7926	0.6536	**0.8622**
LR	0.6939	**0.8102**	0.6117	0.5392
LSTM	0.6217	0.7539	0.5460	0.4797
CC	0.6915	0.8048	0.6124	0.5386
ANN-CC	0.7013	0.7313	0.6807	0.5551
ECC	0.6890	0.8043	0.6072	0.5337
CC*	0.6915	0.8048	0.6124	0.5386
ANN-CC*	0.7036	0.7244	**0.6869**	0.5560

Model	F1 Score	Epochs	Hidden Layers	Training Time (hours)
LSTM [10]	0.7090	200	2	18
ANN-CC	**0.7036**	**100**	**4**	**2**

Table 8.4: Computational Comparison

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 9

Discussion

Our study highlights the critical role of label inter-dependencies in classifying ICD-9 code categories. The inter-dependency within ICD-9 codes is a critical structural feature that drives model performance, and it greatly affects the performance of multi-label classification models in medical text analysis [28].

Our findings reveal that incorporating inter-dependencies using methods like CC and their neural network variant ANN-CC leads to substantial improvements in predictive accuracy. However, the use of an optimal chain order had a relatively minor effect on overall classification performance. This highlights that while modeling inter-dependencies is beneficial, the influence of chain-order heuristics may diminish in more complex or high-dimensional label spaces.

ANN-CC heavily relies on the existence of inter-dependencies in the labels it analyzes [19]. The results of this study also present the potential to uncover more significant and intricate inter-dependent relationships by analyzing a broader set of ICD-9 code categories, which could lead to substantial improvements in classification performance. However, this expansion poses computational challenges and makes it harder to manage cascading errors in methods like CC and ANN-CC, where the order of labels in the chain significantly impacts the performance of the models [21]. Approaches like Ensemble Classifier Chains provide a solution to address these issues, balancing computational feasibility with the ability to fully exploit inter-dependencies.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

The role of label order in improving the performance of CC-based methods further demonstrates the nature of ICD-9 code inter-dependencies. While optimal ordering enhanced results for both CC and ANN-CC in the 10-label scenario, its benefits were marginal when applied to 20 labels, indicating that the more ICD-9 labels are used, the smaller the influence of inter-dependency on the interaction between dataset characteristics and model architecture.

Lastly, the superior performance of the ANN-CC model suggests that neural networks are particularly well-suited for capturing the complex inter-dependencies among ICD-9 codes. However, this advantage is moderated by the increased computational cost and the need for careful model tuning. As demonstrated, while traditional models like Logistic Regression perform adequately without leveraging inter-dependencies, their scalability to larger label sets and unstructured medical texts is limited [14].

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Chapter 10

Conclusion

This study investigated the use of CC integrated with ANNs for multi-label classification of medical text, specifically focusing on automating the assignment of ICD-9 codes using discharge summaries from the MIMIC-III dataset. The aim of this study is to explore the role of label inter-dependencies in improving classification performance.

In conclusion, our study underscores the critical importance of modeling label inter-dependencies in multi-label classification tasks. Our initial assumption has been experimentally verified, since the ANN-CC model not only demonstrated superior predictive performance over traditional methods like Logistic Regression and standalone LSTM but also showcased the benefits of optimizing label order to maximize classification accuracy.

These findings suggest that future research should continue to explore refined chain architectures and optimized label sequencing to further enhance multi-label classification outcomes in healthcare and other domains characterized by complex label structures, while also investigating additional ICD-9 code labels to expand coverage of patient diagnoses and capture richer clinical hierarchies. In particular, refining the ANN architecture for the base classifier of the ANN-CC may offer a more effective means of exploiting the inter-dependencies inherent in such multi-label problems.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Bibliography

[1] Mustafa Ayden, Mehmet Eren Yüksel, and Seniha Yuksel. "A Two-Stream Deep Model for Automated ICD-9 Code Prediction in an Intensive Care Unit". In: *Heliyon* (Feb. 2024). DOI: 10.1016/j.heliyon.2024.e25960.

[2] Everton Alvares Cherman et al. "Lazy multi-label learning algorithms based on mutuality strategies". In: *Journal of Intelligent & Robotic Systems* 80 (2015), pp. 261–276.

[3] Botir Elov, Shahlo Khamroeva, and Zilola Khusainova. "The pipeline processing of NLP". In: *E3S Web of Conferences* 413 (Aug. 2023). DOI: 10.1051/e3sconf/202341303011.

[4] L. Enrique Sucar et al. "Multi-label classification with Bayesian network-based chain classifiers". In: *Pattern Recognition Letters* 41 (2014). Supervised and Unsupervised Classification Techniques and their Applications, pp. 14–22. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2013.11.007. URL: https://www.sciencedirect.com/science/article/pii/S0167865513004303.

[5] Richárd Farkas and György Szarvas. "Automatic construction of rule-based ICD-9-CM coding systems". In: *BMC bioinformatics*. Vol. 9. Springer. 2008, pp. 1–9.

[6] J.C. Ferrão et al. "Using Structured EHR Data and SVM to Support ICD-9-CM Coding". In: *2013 IEEE International Conference on Healthcare Informatics*. 2013, pp. 511–516. DOI: 10.1109/ICHI.2013.79.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

[7] Malte Feucht et al. “Description-based Label Attention Classifier for Explainable ICD-9 Classification”. In: *WNUT*. 2021. URL: https://api.semanticscholar.org/CorpusID:237634985.

[8] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[9] Shuyuan Hu et al. “An explainable CNN approach for medical codes prediction from clinical text”. In: *BMC Medical Informatics and Decision Making* 21 (2021). URL: https://api.semanticscholar.org/CorpusID:231718841.

[10] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. “An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes”. In: *Computer Methods and Programs in Biomedicine* 177 (2019), pp. 141–153. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2019.05.024. URL: https://www.sciencedirect.com/science/article/pii/S0169260718309945.

[11] Alistair Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3 (May 2016), p. 160035. DOI: 10.1038/sdata.2016.35.

[12] Xie Jun et al. “Conditional entropy based classifier chains for multi-label classification”. In: *Neurocomputing* 335 (2019), pp. 185–194. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.01.039. URL: https://www.sciencedirect.com/science/article/pii/S0925231219300591.

[13] Tomasz Kajdanowicz and Przemyslaw Kazienko. “Heuristic Classifier Chains for Multi-label Classification”. In: *Flexible Query Answering Systems*. Ed. by Henrik Legind Larsen et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 555–566. ISBN: 978-3-642-40769-7.

[14] Iuliia Lenivtceva et al. “Applicability of Machine Learning Methods to Multi-label Medical Text Classification”. In: *Computational*

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

Science – ICCS 2020. Ed. by Valeria V. Krzhizhanovskaya et al. Cham: Springer International Publishing, 2020, pp. 509–522. ISBN: 978-3-030-50423-6.

[15] Eneldo Loza Mencía et al. "Tree-based dynamic classifier chains". In: *Machine Learning* 112 (2021), pp. 4129–4165. URL: https://api.semanticscholar.org/CorpusID:245123836.

[16] Neha Nawalkar, Vahida Z. Attar, and Shrida P. Kalamkar. "Automated ICD-9 Medical Code Assignment from Given Free Text Using Deep Learning Approach". In: *Advances in Data and Information Sciences*. Ed. by Shailesh Tiwari et al. Singapore: Springer Singapore, 2022, pp. 317–327. ISBN: 978-981-16-5689-7.

[17] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute. "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques". In: *Journal of the American Medical Informatics Association: JAMIA* 13.5 (2006), pp. 516–525. DOI: 10.1197/jamia.M2077. URL: https://doi.org/10.1197/jamia.M2077.

[18] Aaditya Prakash et al. "Condensed memory networks for clinical diagnostic inferencing". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 3274–3280.

[19] Jesse Read and Jaakko Hollmén. "A Deep Interpretation of Classifier Chains". In: *Advances in Intelligent Data Analysis XIII*. Ed. by Hendrik Blockeel, Matthijs van Leeuwen, and Veronica Vinciotti. Cham: Springer International Publishing, 2014, pp. 251–262. ISBN: 978-3-319-12571-8.

[20] Jesse Read et al. "Classifier Chains for Multi-label Classification". In: *Machine Learning* 85 (Aug. 2009), pp. 254–269. DOI: 10.1007/978-3-642-04174-7_17.

[21] Jesse Read et al. "Classifier chains for multi-label classification". In: *Machine learning* 85 (2011), pp. 333–359.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

[22] Jesse Read et al. “Classifier Chains: A Review and Perspectives”. In: *Journal of Artificial Intelligence Research* 70 (Feb. 2021), pp. 683–718. DOI: 10.1613/jair.1.12376.

[23] Claude Sammut and Geoffrey I. Webb. “Encyclopedia of Machine Learning”. In: *Encyclopedia of Machine Learning*. 2011. URL: https://api.semanticscholar.org/CorpusID:24512455.

[24] Robin Senge, Juan José del Coz, and Eyke Hüllermeier. “On the Problem of Error Propagation in Classifier Chains for Multi-label Classification”. In: *Data Analysis, Machine Learning and Knowledge Discovery*. Ed. by Myra Spiliopoulou, Lars Schmidt-Thieme, and Ruth Janning. Cham: Springer International Publishing, 2014, pp. 163–170. ISBN: 978-3-319-01595-8.

[25] Anandakumar Singaravelan et al. “Predicting ICD-9 Codes Using Self-Report of Patients”. In: *Applied Sciences* 11.21 (2021). ISSN: 2076-3417. DOI: 10.3390/app112110046. URL: https://www.mdpi.com/2076-3417/11/21/10046.

[26] Vishwa Mohan Singh and Omkaresh Kulkarni. “Multi-label classification with classifier chains of ANN models”. In: *2020 IEEE International Conference for Innovation in Technology (INOCON)*. 2020, pp. 1–6. DOI: 10.1109/INOCON50539.2020.9298406.

[27] Sandro Sperandei. “Understanding logistic regression analysis”. In: *Biochemia medica* 24 (Feb. 2014), pp. 12–8. DOI: 10.11613/BM.2014.003.

[28] Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. “Modeling diagnostic label correlation for automatic ICD coding”. In: *arXiv preprint arXiv:2106.12800* (2021).

[29] Grigorios Tsoumakas and Ioannis Katakis. “Multi-Label Classification: An Overview”. In: *International Journal of Data Warehousing and Mining* 3 (Sept. 2009), pp. 1–13. DOI: 10.4018/jdwm.2007070101.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250

[30] Kaushik P Venkatesh, Marium M Raza, and Joseph C Kvedar. "Automating the overburdened clinical coding system: challenges and next steps". In: *npj Digital Medicine* 6.1 (2023), p. 16.

[31] Zhang Yijia et al. "BioWordVec, improving biomedical word embeddings with subword information and MeSH". In: *Scientific Data* 6 (May 2019). DOI: 10.1038/s41597-019-0055-0.

[32] Min Zeng et al. "Automatic ICD-9 coding via deep transfer learning". In: *Neurocomputing* 324 (2019). Deep Learning for Biological/Clinical Data, pp. 43–50. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2018.04.081. URL: https://www.sciencedirect.com/science/article/pii/S0925231218306246.

[33] Min-Ling Zhang and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification". In: *2005 IEEE International Conference on Granular Computing* 2 (2005), 718–721 Vol. 2. URL: https://api.semanticscholar.org/CorpusID:16322186.

Institutional Repository - Library & Information Centre - University of Thessaly
11/09/2026 20:42:37 EEST - 216.73.216.250