sciendo

VOLUME 18, N° 2 2024
Journal of Automation, Mobile Robotics and Intelligent Systems

Diagnostics Based Patient Classification for Clinical Decision Support Systems

Submitted: 28th December 2022; accepted: 13th June 2023

Gaurav Paliwal, Aaquil Bunglowala, Pravesh Kanthed

DOI: 10.14313/JAMRIS/2-2024/16

Abstract:
The widespread adoption of Electronic Healthcare Records has resulted in an abundance of healthcare data. This data holds significant potential for improving healthcare services by providing valuable clinical insights and enhancing clinical decision-making. This paper presents a patient classification methodology that utilizes a multiclass and multilabel diagnostic approach to predict the patient's clinical class. The proposed model effectively handles comorbidities while maintaining a high level of accuracy. The implementation leverages the MIMIC III database as a data source to create a phenotyping dataset and train the models. Various machine learning models are employed in this study. Notably, the natural language processing-based One-Vs-Rest classifier achieves the best classification results, maintaining accuracy and F1 scores even with a large number of classes. The patient diagnostic class prediction model, based on the International Classification of Diseases 9, showcased in this paper, has broad applications in diagnostic support, treatment prediction, clinical assistance, recommender systems, clinical decision support systems, and clinical knowledge discovery engines.

Keywords: *multiclass patient classification, multi-label patient classification, electronic healthcare records, MIMIC III, natural language processing, deep learning*

1. Introduction

Electronic Health Records (EHR) contain a wealth of data about patients, including personal details, medical history, lab reports, diagnostics, and clinical notes from hospital stays [1]. Extracting accurate contextual information and knowledge from EHRs is crucial due to the grouping and linkage of data with previous healthcare events [35]. EHR data comprises both structured and unstructured types. Structured data typically includes personal details and diagnostic/clinical codes, while unstructured data encompasses treatment courses, clinician notes, and lab reports.

Clinical codes in EHRs represent procedures and diagnoses associated with a patient's stay and are used for reporting, management, and billing purposes. They serve as practical sources of information in monitoring and research applications, following standards such as ICPC [5] and ICD [2–4].

The transition from ICD-9 to ICD-10 has significantly increased the complexity of coding systems, making it challenging for caregivers or clinical coders to assign codes. Existing applications have focused mainly on code system browsability but offer limited assistance in coding [6]. Computer-assisted coding systems support caregivers by suggesting codes, providing relevant information, and, in some cases, automatically assigning codes without human intervention [6].

Many clinical coding systems operate in controlled settings, receiving information from limited sources and predicting a restricted set of codes [7]. While some of these approaches perform well, they are difficult to adapt to different environments or scale for larger datasets. Recent research has explored the use of real-world databases, such as MIMIC III, to overcome these limitations [8].

Retrieving information from structured data is relatively easier compared to unstructured data. However, extracting knowledge from structured data alone may yield limited insights and result in inaccuracies and false predictions. Unstructured data, such as clinical notes and lab reports in EHRs, have the potential to provide more accurate knowledge if effectively leveraged. Manually labeling a vast volume of unstructured data from various sources is time-consuming and impractical. Automated information retrieval using Natural Language Processing (NLP) offers a more efficient and scalable alternative to manual methods [36].

Identifying groups of individuals with shared characteristics can greatly benefit the healthcare industry. These findings can be utilized in various secondary analyses, including computational phenotyping for patient cohort identification, clinical decision support, evidence-based treatment, and research. This paper presents the implementation of multiclass patient classification using different machine and deep learning models on structured data. The models utilize libraries such as sklearn, Keras, and TensorFlow APIs, with MIMIC III serving as the data source. The system automatically assigns codes to patients based on clinical diagnoses. The paper also addresses unstructured data by performing multi-label classification using NLP models, specifically employing a one-vs-rest classifier.

The paper is organized into six sections. Sections 1 and 2 provide an introduction and literature review, respectively. Section 3 describes the design methodology employed, while Section 4 focuses on implementation details. Results and conclusions are presented in Sections 5 and 6, respectively.

2. Literature Review

Machine learning and deep learning have gained significant attention in recent years, leading to extensive research in this field. However, most of the work in automated clinical code generation has been conducted on limited dummy data [9, 10, 39, 62]. To address this, recent techniques have been developed for layered prediction and classification of real-world datasets, leveraging the sparsity of output codes [9, 10, 39]. The MIMIC database is commonly used as a benchmark by researchers to evaluate new technologies [8]. For instance, Perotte et al utilized the MIMIC II database to predict higher-level diagnostic codes, which was further extended to predict complete codes [9]. Other researchers have focused on predicting ICD 10 PCS codes from clinical discharge summaries [10].

While these methodologies depend on accessible and comprehensive clinical discharge reports, many clinical summaries still lack important information, and different service providers and hospitals may have varying styles of maintaining details. To compensate for missing information, techniques have been proposed for predicting and identifying patient cohorts based on high-throughput phenotyping [3,11, 12,40].

A unified data view has been achieved by mapping structured and unstructured data to standardized UMLS [12, 13, 37, 38]. Although this approach is powerful for data conversion, it introduces dependencies on the used ontologies, requiring substantial efforts to map local ontologies in clinical notes or terms [14,41]. Unstructured data sources can also be converted or mapped using a bag-of-words representation, either through early or late data integration from the source itself [15]. The density of information features from different sources can create stronger or weaker structured representations. Directly feeding data from the source to the meta-classifier often yields better results, but this may result in data loss due to a single data point [60].

A recent study introduced a deep neural network with stacked auto-encoders, known as “deep patient,” to create vector representations of patients [16]. The structured and unstructured data sources were preprocessed using subject modeling for data generalization [17]. This data representation method can be applied to diverse medical applications, focusing only on the most interesting features [42,61].

For limited data, algorithmic performance can be improved using feature selection algorithms [18, 19, 43]. In classification tasks, various approaches are available to improve fitness by ranking features towards a class. One such method aims to find a nominal set of features by reducing redundancies among them [20]. This approach selects features strongly correlated with optimizing the classification task. However, relying solely on strongly correlated features may not contribute to accurate predictions [21].

While feature-based redundancy calculation methods can address these issues, they may not yield significant results when applied to a large number of classes or features [22]. Another feature selection technique, the Markov blanket model, offers good results for feature selection. This method utilizes a Bayesian network and provides efficient solutions.

Noteworthy patient classification techniques found during the survey include disease phenotypes in EHRs using machine learning [28], an algorithm to recognize patients with Autism Spectrum Disorder [29], and learning from heterogeneous temporal data [30].

Additionally, “Interpretable patient classification using integrated patient similarity networks” presents a novel approach to classification using clinical predictors based on genomic data [31]. Pierre Courtio et al. demonstrated the use of patient classification for predicting outcomes in “Deep learning-based classification of mesothelioma improves prediction of patient outcome” [32].

3. Methodology

In this study, the MIMIC-III database was utilized as the primary source of data. The MIMIC-III, short for Medical Information Mart for Intensive Care III, is a publicly available clinical database [25]. It contains comprehensive information about cancer patients who were hospitalized in the critical care unit at the Beth Israel Deaconess Medical Center between 2001 and 2012. The MIMIC-III database encompasses healthcare records of approximately 40,000 patients, including their basic demographic information, diagnostics, vital signs, clinical procedures, laboratory test results, caregiver notes, medications, and mortality data. This extensive dataset in MIMIC-III supports various types of studies, ranging from the development of decision support systems to electronic tool advancements in healthcare.

For this particular study, we utilized the MIMIC-III database as an electronic healthcare record to classify cancer patients based on the provided diagnostics. This classification can be further employed to assist doctors in predicting the most appropriate course of action for each patient.

The MIMIC-III database comprises 40 tables with 534 columns and nearly 720 million records. These tables and records are interconnected using identifiers such as subject ID and admission ID. It is worth noting that all data within the database have been fully de-identified to ensure the preservation of patient privacy and confidentiality.

Table	Children	Parents	Columns	Rows	Comments
admissions	18	1	19	58,976	Hospital admissions associated with an ICU stay.
callout		2	24	34,499	Record of when patients were ready for discharge (called out), and the actual time of their discharge (or more generally, their outcome).
caregivers	7		4	7,567	List of caregivers associated with an ICU stay.
chartevents		5	15	330,712,483	Events occuring on a patient chart.
chartevents_1			15	38,033,561	Partition of chartevents. Should not be directly queried.
chartevents_10			15	9,584,888	Partition of chartevents. Should not be directly queried.
chartevents_11			15	470,141	Partition of chartevents. Should not be directly queried.
chartevents_12			15	265,413	Partition of chartevents. Should not be directly queried.
chartevents_13			15	39,066,570	Partition of chartevents. Should not be directly queried.
chartevents_14			15	100,075,138	Partition of chartevents. Should not be directly queried.
chartevents_2			15	13,116,197	Partition of chartevents. Should not be directly queried.
chartevents_3			15	38,657,533	Partition of chartevents. Should not be directly queried.
chartevents_4			15	9,374,587	Partition of chartevents. Should not be directly queried.
chartevents_5			15	18,201,026	Partition of chartevents. Should not be directly queried.
chartevents_6			15	28,014,688	Partition of chartevents. Should not be directly queried.
chartevents_7			15	255,967	Partition of chartevents. Should not be directly queried.
chartevents_8			15	34,322,082	Partition of chartevents. Should not be directly queried.
chartevents_9			15	1,274,692	Partition of chartevents. Should not be directly queried.
cptevents		2	12	573,146	Events recorded in Current Procedural Terminology.
d_cpt			9	134	High-level dictionary of the Current Procedural Terminology.
d_icd_diagnoses	1		4	14,710	Dictionary of the International Classification of Diseases, 9th Revision (Diagnoses).
d_icd_procedures	1		4	3,898	Dictionary of the International Classification of Diseases, 9th Revision (Procedures).
d_items	8		10	12,487	Dictionary of non-laboratory-related charted items.
d_labitems	1		6	753	Dictionary of laboratory-related items.
datetimeevents		5	14	4,485,937	Events relating to a datetime.
diagnoses_icd		3	5	651,047	Diagnoses relating to a hospital admission coded using the ICD9 system.
drgcodes		2	8	125,557	Hospital stays classified using the Diagnosis-Related Group system.
icustays	8	2	12	61,532	List of ICU admissions.
inputevents_cv		4	22	17,527,935	Events relating to fluid input for patients whose data was originally stored in the CareVue database.
inputevents_mv		5	31	3,618,991	Events relating to fluid input for patients whose data was originally stored in the MetaVision database.
labevents		3	9	27,854,055	Events relating to laboratory tests.
microbiologyevents		5	16	631,726	Events relating to microbiology tests.
noteevents		3	11	2,083,180	Notes associated with hospital stays.
outputevents		5	13	4,349,218	Outputs recorded during the ICU stay.
patients	19		8	46,520	Patients associated with an admission to the ICU.
prescriptions		3	19	4,156,450	Medicines prescribed.
procedureevents_mv		5	25	258,066	Procedure start and stop times recorded for MetaVision patients.
procedures_icd		3	5	240,095	Procedures relating to a hospital admission coded using the ICD9 system.
services		2	6	73,343	Hospital services that patients were under during their hospital stay.
transfers		3	13	261,897	Location of patients during their hospital stay.
40 Tables			**534**	**728,556,685**	

Figure 1. MIMIC III schema details

Due to the enormous size of the data in the MIMIC-III database, interpreting and understanding the data posed a significant challenge. The database contained data distributed across various tables, as illustrated in Figure 1. To facilitate data analysis, the MIMIC-III database was initially imported into PostgreSQL. An interface to the database was developed to gain a comprehensive understanding of the MIMIC-III schema. Figures 2 and 3 depict the interface designed to interact with the MIMIC-III database.

For a visual demonstration of the graphical user interface (GUI), a video showcasing its functionality is available on YouTube via the following link: https://www.youtube.com/watch?v=9eNtfb3oR-E&lc=UgyHko8pYeyGp9H-DHp4AaABAg. Additionally, the front-end design code for the interface can be downloaded from GitHub using the following link: https://github.com/gvpaliwal/Mimic-III.

For analysis and prediction, the data primarily selected from the MIMIC III dataset was obtained from five main tables: admissions, drgcodes, diagnoses_icd, d_icd_diagnoses, and noteevents. These tables provided essential information for the study, including patient admissions details, disease diagnostics in ICD9 codes, and clinical notes.

To address the issue of data imbalance, various data preprocessing and selection techniques were employed. The first four tables, namely admissions, drgcodes, diagnoses_icd, and d_icd_diagnoses, provided structured data that could be preprocessed and utilized. These tables contained patient details, disease diagnostics in the form of ICD9 codes, and the diagnosis code associated with the disease for which the patient was admitted or billed. By considering the diagnosis code as the class label for the patient, the dataset enabled the application of both supervised and unsupervised machine learning techniques.

On the other hand, the noteevents table contained unstructured text data in the form of clinical notes. This text data was suitable for applying natural language processing (NLP) models. The main focus of the study was to assign single or multiple class labels to a patient based on the diagnostics and clinical notes.

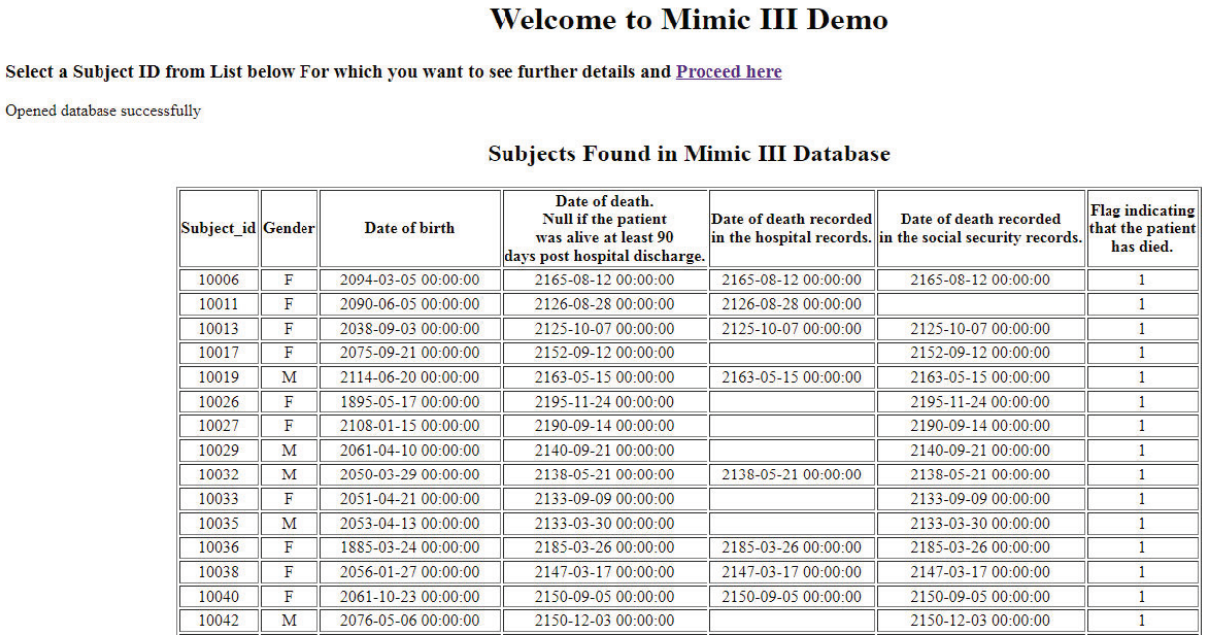
The study aimed to develop a comprehensive approach to patient classification and prediction by combining the structured and unstructured data from these tables.

4. Implementation Details

Machine Learning in healthcare has demonstrated its power in sorting and classifying health data, as well as accelerating doctors' clinical decision-making process by providing predictions that can save lives and simplify tasks. For the task of Patient Classification/Cohort Identification, our primary focus was on utilizing ML/DL models that have shown promising accuracy when applied to healthcare data for similar tasks.

In this section, we will present implementations of multiclass patient classification using various machine and deep learning models specifically designed for structured data. These models aim to assign codes to patients based on their clinical diagnosis automatically. Furthermore, we will delve into the realm of unstructured data and employ Natural Language Processing (NLP) models to perform multi-label classification. This approach enables us to extract meaningful information from clinical notes and make predictions based on the unstructured data.

By combining the strengths of both structured and unstructured data analysis, we aim to develop a comprehensive framework for patient classification and prediction in healthcare settings.



Welcome to Mimic III Demo

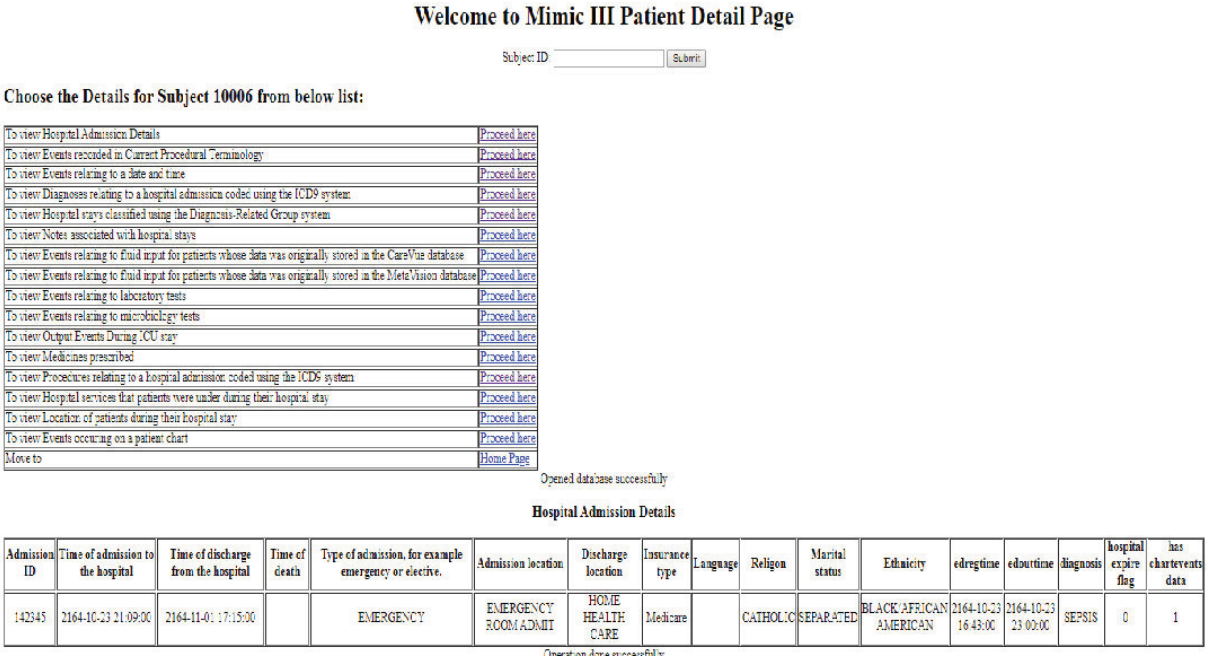
Select a Subject ID from List below For which you want to see further details and Proceed here

Opened database successfully

Subjects Found in Mimic III Database

Subject_id	Gender	Date of birth	Date of death. Null if the patient was alive at least 90 days post hospital discharge.	Date of death recorded in the hospital records.	Date of death recorded in the social security records.	Flag indicating that the patient has died.
10006	F	2094-03-05 00:00:00	2165-08-12 00:00:00	2165-08-12 00:00:00	2165-08-12 00:00:00	1
10011	F	2090-06-05 00:00:00	2126-08-28 00:00:00	2126-08-28 00:00:00		1
10013	F	2038-09-03 00:00:00	2125-10-07 00:00:00	2125-10-07 00:00:00	2125-10-07 00:00:00	1
10017	F	2075-09-21 00:00:00	2152-09-12 00:00:00		2152-09-12 00:00:00	1
10019	M	2114-06-20 00:00:00	2163-05-15 00:00:00	2163-05-15 00:00:00	2163-05-15 00:00:00	1
10026	F	1895-05-17 00:00:00	2195-11-24 00:00:00		2195-11-24 00:00:00	1
10027	F	2108-01-15 00:00:00	2190-09-14 00:00:00		2190-09-14 00:00:00	1
10029	M	2061-04-10 00:00:00	2140-09-21 00:00:00		2140-09-21 00:00:00	1
10032	M	2050-03-29 00:00:00	2138-05-21 00:00:00	2138-05-21 00:00:00	2138-05-21 00:00:00	1
10033	F	2051-04-21 00:00:00	2133-09-09 00:00:00		2133-09-09 00:00:00	1
10035	M	2053-04-13 00:00:00	2133-03-30 00:00:00		2133-03-30 00:00:00	1
10036	F	1885-03-24 00:00:00	2185-03-26 00:00:00	2185-03-26 00:00:00	2185-03-26 00:00:00	1
10038	F	2056-01-27 00:00:00	2147-03-17 00:00:00	2147-03-17 00:00:00	2147-03-17 00:00:00	1
10040	F	2061-10-23 00:00:00	2150-09-05 00:00:00	2150-09-05 00:00:00	2150-09-05 00:00:00	1
10042	M	2076-05-06 00:00:00	2150-12-03 00:00:00		2150-12-03 00:00:00	1

Figure 2. MIMIC III database front end



Welcome to Mimic III Patient Detail Page

Subject ID Submit

Choose the Details for Subject 10006 from below list:

To view Hospital Admission Details	Proceed here
To view Events recorded in Current Procedural Terminology	Proceed here
To view Events relating to a date and time	Proceed here
To view Diagnoses relating to a hospital admission coded using the ICD9 system	Proceed here
To view Hospital stays classified using the Diagnosis-Related Group system	Proceed here
To view Notes associated with hospital stays	Proceed here
To view Events relating to fluid input for patients whose data was originally stored in the CareVue database	Proceed here
To view Events relating to fluid input for patients whose data was originally stored in the MetaVision database	Proceed here
To view Events relating to laboratory tests	Proceed here
To view Events relating to microbiology tests	Proceed here
To view Output Events During ICU stay	Proceed here
To view Medicines prescribed	Proceed here
To view Procedures relating to a hospital admission coded using the ICD9 system	Proceed here
To view Hospital services that patients were under during their hospital stay	Proceed here
To view Location of patients during their hospital stay	Proceed here
To view Events occuring on a patient chart	Proceed here
Move to	Home Page

Opened database successfully

Hospital Admission Details

Admission ID	Time of admission to the hospital	Time of discharge from the hospital	Time of death	Type of admission, for example emergency or elective.	Admission location	Discharge location	Insurance type	Language	Religion	Marital status	Ethnicity	edregtime	edouttime	diagnosis	hospital expire flag	has chartevents data
142345	2164-10-23 21:09:00	2164-11-01 17:15:00		EMERGENCY	EMERGENCY ROOM ADMIT	HOME HEALTH CARE	Medicare		CATHOLIC	SEPARATED	BLACK/AFRICAN AMERICAN	2164-10-23 16:43:00	2164-10-23 23:00:00	SEPSIS	0	1

Figure 3. MIMIC III database front end

4.1. Data Preprocessing

In the initial dataset extracted from the MIMIC III database, there were 651,047 records corresponding to 58,976 admissions and 46,520 unique patient stays. The dataset included a total of 6,985 unique ICD-9 diagnostic codes. To simplify the analysis, these diagnostic codes were aggregated into 891 ICD-9 diagnostic block chapter codes based on the descriptions provided in the database. These diagnostic block chapter codes were then used to predict the class or assign labels to the patients.

As a preprocessing step, we removed the diagnostic block chapter codes related to external injuries, poisoning, or any supplementary classifications, as these causes are event-driven and not suitable for analytics. After this removal, we were left with a total of 625 ICD-9 diagnostic block chapter codes.

Due to the large number of classes, accurately determining the class using any predictive model became challenging. Additionally, the data remained skewed and unbalanced due to the varying frequencies of different diseases. To address this issue, we selected specific datasets from hospital stays based on the occurrence frequencies of the ICD-9 block chapter

codes. This approach allowed us to create balanced datasets for training and evaluation purposes.

Table 1 summarizes the occurrence frequencies of different ICD-9 diagnostic codes, which influenced the selection of datasets for analysis.

After applying the designated thresholds to select hospital stays with different frequent chapter codes, we obtained groups of hospital stays with varying frequencies. The most frequent chapter code had an occurrence frequency of 21,329. We then linked the HCFA drug codes from the drgcodes table to the selected hospital stays, which represented the drugs for which the patients were treated and billed.

The noteevents table in the MIMIC III database contained a wealth of unstructured text information, including various clinical notes such as demographic details, services provided, allergies, chief complaint, medical history, social history, physical exam findings, pathology and scan reports, medication details, discharge information, and follow-up instructions. To clean the data, we removed stop words and special characters and converted the text to lowercase, preparing it for further analysis.

After the data preprocessing steps, we ended up with four dataset groups, each containing records with specific ICD-9 chapter frequencies. These groups were suitable for various predictive analytics tasks. The next step was to split the data into train and test sets and apply different learning models for analysis and prediction purposes.

4.2. Model Implementation on Structured Data

4.2.1. K-nearest Neighbor

The k-nearest neighbors (KNN) algorithm is a supervised classification method that utilises proximity to make predictions about a given data point. It can be applied for both classification and regression tasks. In our study, we employed the KNN algorithm as a classification model to predict and identify patients with similar characteristics based on their diagnostic information [44–46].

To implement the KNN algorithm, we utilized the KNeighborsClassifier and RadiusNeighborsClassifier with a specified number of neighbors set to 25 for KNeighborsClassifier and the default radius for RadiusNeighborsClassifier. The distance weight function was used in the prediction process. The algorithm parameter was set to "auto", allowing the model to automatically select the most suitable algorithm for computing the nearest neighbors based on the provided values. We experimented with different distance metrics for computation but ultimately found that the default metric, Minkowski, yielded the best results, so we utilised it in our implementation.

4.2.2. Naive Bayes Classifier

The Naive Bayes (NB) algorithm is a classification method that assumes independence between features and utilizes probability theory to make predictions. It has been widely used in various medical applications and has shown good performance [59]. Although the independence assumptions may not always hold true in reality, Naive Bayes can still provide effective results in many medical scenarios [47, 50].

In our study, we trained the Naive Bayes algorithm in a supervised learning setting. We implemented three different variants of the NB classifier: GaussianNB, CategoricalNB, and MultinomialNB. The prior probabilities in the model were set to none, and the variance smoothing was set to the default value. Prior to feeding the data into the Naive Bayes classifier, we applied min-max scaling to normalize the data. The additive smoothing parameter, also known as the alpha value, was set to 1 in our implementation.

4.2.3. Decision Tree Classifier

The decision tree algorithm is a supervised learning approach that can be used for both regression and classification tasks. It employs a greedy search strategy to find the best split points in a tree-like structure. The decision tree algorithm follows a divide and conquer methodology, recursively splitting the data until most or all records are classified under different labels [49].

In our study, we utilised the decision tree algorithm along with cross-validation to evaluate how well the model fits the data. Additionally, we explored the random forest method, which improves accuracy by combining multiple decision trees. Random forest classifiers are particularly effective when the individual trees are uncorrelated [48]. Decision trees are highly interpretable, making them popular in healthcare analytics (ch8dt). They can handle complex data patterns and serve as the foundation for ensemble models [51, 52].

We implemented several decision tree models using scikit-learn, including Decision Tree Classifier, AdaBoost Classifier, Random Forest Classifier, Bagging Classifier, and Gradient Boosting Classifier. The number of trees or estimators in the random forest was set to 100. Due to computational limitations, we set the maximum depth of the trees to 15. The quality of a split was measured using the Gini criterion, and the remaining parameters were kept as default or set to "auto".

4.2.4. Support Vector Machines

Support Vector Machines (SVMs) are a reliable classification method that can increase a model's predictive accuracy without overfitting the training set. They are particularly effective when dealing with datasets that have many predictors. In the context of electronic health record (EHR) data, SVM-based methods have been utilised in the literature to address classification tasks and have been shown to provide quick, accurate, and reliable results, especially when dealing with imbalanced classes [53, 54].

In our study, we implemented the support vector classifier using different kernels, including linear, polynomial, and radial basis function (rbf) kernels. The regularization parameter (C) was set to 1, which controls the trade-off between achieving a low training error and allowing for a larger margin.

Table 1. ICD-9 Diagnostic codes occurrence frequency

Sr. No.	Item	Occurrence frequency	Number of Items
1	ICD-9 diagnostic codes	>1000	98
2		1000 to 100	557
3		<100	3970
	ICD-9 block Chapter code	>15000	6
4		>1000	109
5		1000 to 100	201
6		<100	315

For the polynomial kernel, the degree of the kernel function was limited to 5. Other parameters such as the kernel coefficient, kernel function independent term, shrinking heuristic, probability estimates, and tolerance values were set to their default or "auto" values. These settings were chosen to provide a balanced and effective classification model.

4.2.5. Deep Neural Network

Artificial Neural Networks (ANNs) are a subset of machine learning techniques that can be powerful tools in healthcare diagnosis and analytics, leading to improved healthcare delivery and reduced costs [55,56]. In our study, we employed a multiclass neural network classifier to categorise and group data effectively.

Implementing multiclass neural networks can be challenging due to the nonlinear nature of class labels. To address this, we represented the labels as binary vectors of length K, where a 1 at the kth position corresponds to a label of k. An activation function was applied at the output to learn the output vector. This approach allowed us to generate valid probability distributions by summing over all the values of k to normalize the likelihood [26].

We used a traditional training methodology to train the model but with an objective gradient specifically designed for multinomial logistic regression [33, 34]. We experimented with various combinations of parameters to fine-tune the model for better results.

The tuned deep learning neural network architecture utilised a baseline sequential model with dense layers. The model employed the Rectified Linear Unit (relu) activation function for hidden layers and the softmax activation function for the output layer. Categorical cross-entropy was used as the loss function to mitigate training losses, with the Adam optimiser for model optimisation. The evaluation metric used was accuracy. The model was trained using 10-fold cross-validation with 200 epochs.

To assess the performance of the classifier model, we calculated the baseline accuracy, which serves as a benchmark for comparison.

4.2.6. NLP-based Models

Natural Language Processing (NLP) is a field of study that enables computers to understand and interpret human language, both spoken and written.

By combining machine learning, statistics, deep learning models, and rule-based computational linguistics modelling, NLP allows computers to process and analyse large volumes of narrative information, such as clinical notes in healthcare systems. This is particularly important because clinical notes are often stored in non-standardized and unstructured formats, making it challenging to extract meaningful insights from them.

In our implementation, we utilised NLP techniques to capture and analyse unstructured information from clinical notes. We employed the MultiLabelBinarizer from the sklearn library to transform the clinical notes into target variables. These target variables were then used to calculate the term frequency-inverse document frequency (tf-idf) vector values, which represent the importance of words in the clinical notes. We set a maximum of 10,000 features for the tf-idf representation.

For classification, we used the OneVsRestClassifier, which allows us to handle multi-label classification tasks. This classifier trains multiple binary classifiers, each treating one label as the positive class and the rest as the negative class. We evaluated the model's performance using the F1 score, which is a measure of the model's accuracy in classifying multiple labels.

By employing NLP techniques, we aimed to transform the unstructured information in clinical notes into a format that can be understood and utilised by a decision support system, enabling more effective healthcare analytics and decision-making.

5. Results and Discussion

The machine learning and NLP implementations are done on 4 frequency-based selected datasets. The Naive Bayes classifiers' performance was weakest among the implemented models on the given data set. The deep learning and decision trees models performed reasonably well with fewer classes but failed to maintain accuracy when the number of classes was increased. Table 2 shows the detailed performance measure results for different models.

The NLP-based model provided the best scores in diagnostic label prediction. The information loss while training the model was calculated as 0.005. The true positive rate for the model was calculated to be around 99%. Figures 4 and 5 can be referred to for the comparative analysis of accuracies and F1 scores of different models, respectively. Figure 6 shows a prediction screenshot of an NLP-based multi-label classifier.

Table 2. Model performance measure results

Sr. N.	Learning Model	Implemented Variant	Frequency of chapter codes	No. of Chapter codes	No. of records	Performance Measure
1	K-Nearest neighbor	K Neighbors Classifier	>15000	6	42165	Accuracy: 96.44 , F1 score: 0.93
			>1000	109	54903	Accuracy: 73.21 , F1 score: 0.62
			1000 to 100	310	56059	Accuracy: 23.66 , F1 score: 0.42
			<100	625	56202	Accuracy: 10.01 , F1 score: 0.21
		Radius Neighbors Classifier	>15000	6	42165	Accuracy: 97.31, F1 score: 0.94
			>1000	109	54903	Accuracy: 78.56, F1 score: 0.68
			1000 to 100	310	56059	Accuracy: 26.41, F1 score: 0.41
			<100	625	56202	Accuracy: 18.66 , F1 score: 0.22
2	Naive Bayes classifier	Gaussian NB	>15000	6	42165	Accuracy: 76.66 , F1 score: 0.65
			>1000	109	54903	Accuracy: 26.48 , F1 score: 0.51
			1000 to 100	310	56059	Accuracy: 11.25 , F1 score: 0.23
			<100	625	56202	Accuracy: 9.07, F1 score: 0.18
		Categorical NB	>15000	6	42165	Accuracy: 78.22, F1 score: 0.69
			>1000	109	54903	Accuracy: 43.55, F1 score: 0.44
			1000 to 100	310	56059	Accuracy: 18.64, F1 score: 0.26
			<100	625	56202	Accuracy: 16.11, F1 score: 0.22
		Multinomial NB	>15000	6	42165	Accuracy: 79.21, F1 score: 0.72
			>1000	109	54903	Accuracy: 44.00, F1 score: 0.43
			1000 to 100	310	56059	Accuracy: 17.88, F1 score: 0.25
			<100	625	56202	Accuracy: 16.01 , F1 score: 0.22

Table 2. Continued

Sr. N.	Learning Model	Implemented Variant	Frequency of chapter codes	No. of Chapter codes	No. of records	Performance Measure
3	Decision Tree classifier	Decision Tree Classifier	>15000	6	42165	Accuracy: 78.42, F1 score: 0.80
			>1000	109	54903	Accuracy: 25.48, F1 score: 0.51
			1000 to 100	310	56059	Accuracy: 13.45, F1 score: 0.13
			<100	625	56202	Accuracy: 9.60%, F1 score: 0.12
		Ada Boost Classifier	>15000	6	42165	Accuracy: 78.22, F1 score: 0.80
			>1000	109	54903	Accuracy: 23.44, F1 score: 0.48
			1000 to 100	310	56059	Accuracy: 12.56, F1 score: 0.21
			<100	625	56202	Accuracy: 10.23, F1 score: 0.20
		Random Forest Classifier	>15000	6	42165	Accuracy: 98.64, F1 score: 0.96
			>1000	109	54903	Accuracy: 86.55, F1 score: 0.91
			1000 to 100	310	56059	Accuracy: 73.23, F1 score: 0.86
			<100	625	56202	Accuracy: 25.60, F1 score: 0.59
		Bagging Classifier	>15000	6	42165	Accuracy: 96.75, F1 score: 0.97
			>1000	109	54903	Accuracy: 86.49, F1 score: 0.89
			1000 to 100	310	56059	Accuracy: 72.59, F1 score: 0.72
			<100	625	56202	Accuracy: 24.23, F1 score: 0.62
		Gradient Boosting Classifier	>15000	6	42165	Accuracy: 96.87, F1 score: 0.97
			>1000	109	54903	Accuracy: 87.54, F1 score: 0.90
			1000 to 100	310	56059	Accuracy: 70.27, F1 score: 0.73
			<100	625	56202	Accuracy: 22.45, F1 score: 0.64

Table 2. Continued

4	Support Vector Machines	linear kernel	>15000	6	42165	Accuracy: 64.23, F1 score: 0.51
			>1000	109	54903	Accuracy: 16.21, F1 score: 0.23
			1000 to 100	310	56059	Accuracy: 11.12, F1 score: 0.21
			<100	625	56202	Accuracy: 9.54, F1 score: 0.19
		polynomial kernel	>15000	6	42165	Accuracy: 45.21, F1 score: 0.41
			>1000	109	54903	Accuracy: 12.87, F1 score: 0.15
			1000 to 100	310	56059	Accuracy: 6.23, F1 score: 0.12
			<100	625	56202	Accuracy: 3.06, F1 score: 0.11
		radial basis function kernel	>15000	6	42165	Accuracy: 63.68, F1 score: 0.50
			>1000	109	54903	Accuracy: 17.11, F1 score: 0.24
			1000 to 100	310	56059	Accuracy: 11.21, F1 score: 0.21
			<100	625	56202	Accuracy: 9.64, F1 score: 0.19
5	Deep Neural Network	Artificial neural network	>15000	6	42165	Accuracy: 98.86, F1 score: 0.96
			>1000	109	54903	Accuracy: 83.45, F1 score: 0.92
			1000 to 100	310	56059	Accuracy: 74.03 , F1 score: 0.84
			<100	625	56202	Accuracy: 63.54, F1 score: 0.76
6	NLP based Logistic Regression	One Vs Rest Classifier	>15000	6	42165	Accuracy: 98.68, F1 score: 0.98, TP: 100
			>1000	109	54903	Accuracy: 87.23, F1 score: 0.96, TP: 99.6
			1000 to 100	310	56059	Accuracy: 86.42, F1 score: 0.92, TP: 99.4
			<100	625	56202	Accuracy: 82.86, F1 score: 0.89, TP: 98.8

To analyze the relation between accuracy and records, a less formal approach was used to measure the strength of the relationship between them. Spearman's rank correlation coefficient testing is applied to test the hypothesis of association. The pairs of observations have been randomly selected and ranks are assigned within the sample.

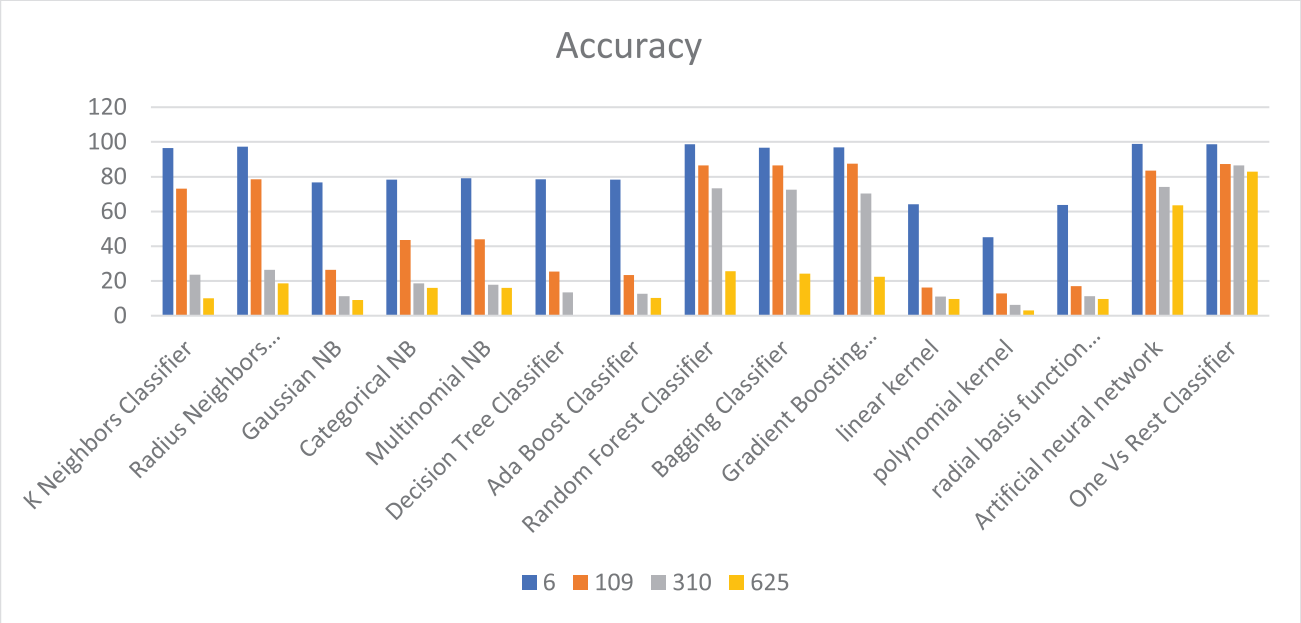


Figure 4. Different machine learning model accuracies

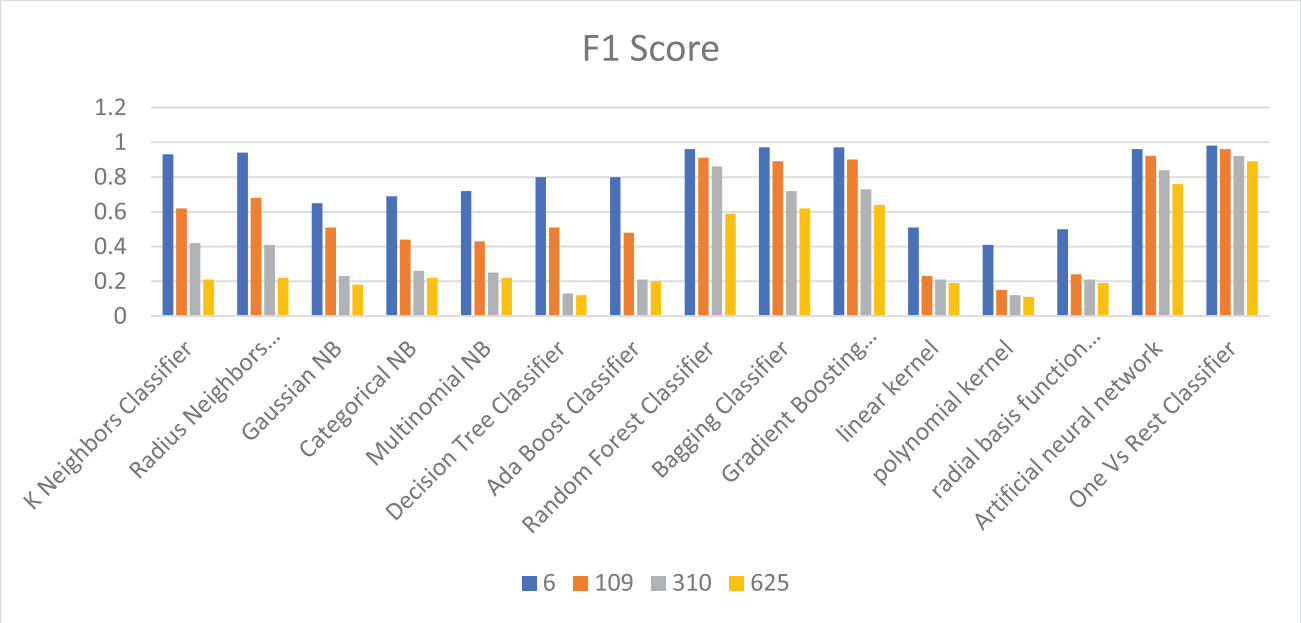


Figure 5. Different machine learning model F1 scores

```
for i in range(5):
    k = xval.sample(1).index[0]
    print("HADM_ID: ", NOTE_DIAGNOSES['HADM_ID'][k], infer_tags(xval[k])), print("Actual ICDs: ",NOTE_DIAGNOSES['TARGET'][k], "\n")
```

```
HADM_ID:  108453.0 [('041', '049', '401', '599', '682')]
Actual ICDs:  ['518', '599', '682', '049', '584', '707', '428', '276', '041', '285', '401', '272', '414', '041', '296', '276']

HADM_ID:  167274.0 [('038', '256', '276', '403', '427', '428', '458', '518', '585', '785')]
Actual ICDs:  ['038', '585', '403', '785', '286', '453', '578', '440', '276', '250', '414', '280', '593', '553', '715', '733', '424']

HADM_ID:  127551.0 [('076', '765', '770', '774', '779')]
Actual ICDs:  ['765', '076', '774', '765', '770', '779']

HADM_ID:  105875.0 [('765', '774')]
Actual ICDs:  ['765', '774', '765']

HADM_ID:  112372.0 [('584',)]
Actual ICDs:  ['576', '572', '584', '276', '428', '427', '272', '428']
```

Figure 6. NLP based multi-label classification

Table 3. t statistics analysis details for spearman's rank correlation hypothesis testing

Parameter	Value
Correlation Coefficient (rs)	0.983884884
Sample size (n)	10
degree of freedom (df) = (n−2)	8
t statistic	15.56375323
level of Significance (α)	0.05
t critical value	2.306004135
p value	2.89393E-07

This test makes no assumptions about the probability relation between two variables. H_o: *there is no relationship between accuracy and records.* H_1: *there is a relationship between accuracy and records.* The details of the analysis are shown in Table 3.

As per the results obtained from spearman's rank correlation hypothesis testing the p value is less than the alpha value. Hence, H_0 is rejected i.e. there is a relationship found between accuracy and records. Using the test result, we can conclude that there exists a relationship between accuracy and records prediction on the given data considering the number of classes to be constant.

6. Conclusion and Future Direction

The presented patient classification model holds great potential for the healthcare system. The superior performance of the NLP-based multi-label prediction models suggests their effectiveness in accurately predicting a patient's class based on the provided diagnostics. With a larger number of records available for training the model, it is expected that the prediction accuracies of these models will improve even further.

This model's applications are wide-ranging and can greatly benefit various aspects of the healthcare system. Some potential applications include diagnostic support, line of treatment prediction, clinical assistance, recommender systems, clinical decision support systems, and clinical knowledge discovery engines. These predictions can provide valuable insights and assist healthcare professionals in making informed decisions, thereby improving the quality and efficiency of patient care.

More advanced NLP-based methods can be explored and implemented to enhance the predictions' performance. These advancements may include incorporating techniques such as transfer learning, contextual embeddings, or attention mechanisms, which have shown promising results in natural language processing tasks.

The statistical test applied to assess the model accuracy clearly indicates that as the number of records increases, the accuracy of the model is expected to improve. This suggests that collecting and incorporating more patient data into the training process will likely yield more accurate predictions.

It is worth noting that the data used for training the model, particularly when dealing with a larger number of diagnostic chapter codes, may be skewed due to uneven frequency of disease occurrences. However, the NLP-based models demonstrated the ability to maintain accuracy and F1 scores, indicating their robustness in handling such imbalances.

In the future, incorporating additional details available in the MIMIC III database can further enhance the prediction accuracies. These additional details may include patient demographics, medical history, laboratory results, medications, and other relevant factors that can provide a more comprehensive understanding of the patients' conditions and facilitate more accurate predictions.

Overall, the presented patient classification model, particularly the NLP-based approaches, shows great potential for improving decision support and point-of-care service delivery in healthcare systems. Further advancements and incorporation of richer patient data can lead to even more accurate and valuable predictions.

AUTHORS

Gaurav Paliwal* – SVKM's, Narsee Monjee Institute of Management Studies, Indore, India, e-mail: gaurav.paliwal@nmims.edu.

Aaquil Bunglowala – Sri Aurobindo Institute of Technology, Indore, India, e-mail: aaquilbun@gmail.com.

Pravesh Kanthed – Choithram Interventional Spine & Pain Centre, Indore, India, e-mail: praveshkanthed@gmail.com.

*Corresponding author

References

[1] C.-J. Hsiao, E. Hing, "Use and characteristics of electronic health record systems among office-based physician practices," *NCHS Data Brief*, vol. 111, 2012, pp. 1–8.

[2] G.S. Alotaibi, C. Wu, A. Senthilselvan, M.S. McMurtry, "The validity of ICD codes coupled with imaging procedure codes for identifying acute venous thromboembolism using administrative data," *Vasc. Med.*, vol. 20, no. 4, 2015, pp. 364–368. doi: 10.1177/1358863X15573839.

[3] W.Q. Wei, P.L. Teixeira, H. Mo, R.M. Cronin, J.L. Warner, J.C. Denny, "Combining billing codes, clinica notes, and medications from electronic health records provides superior phenotyping performance," *J. Am. Med. Inform. Assoc.*, vol. 23 , no. e1, 2016, pp. 20–27. doi: 10.1093/jamia/ocv130.

[4] World Health Organization, "International Classification of Diseases (icd)," 2012.

[5] H. Lamberts, I. Okkes, et al, "Icpc-2," *International Classification of Primary Care*, 1998.

[6] S. Pakhomov, J.D. Buntrock, C.G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *J. Am. Med. Inform. Assoc.*, vol. 13, no. 5, 2006, pp. 516–525, doi: 10.1197/jamia.M2077.

[7] M.H. Stanfill, M. Williams, S.H. Fenton, R.A. Jenders, W.R. Hersh, "A systematic literature review of automated clinical coding and classification systems," *J. Am. Med. Inform. Assoc.*, vol. 17, no. 6, 2010, pp. 646–651, doi: 10.1136/jamia.2009.001024.

[8] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, "A freely accessible critical care database," *Sci. Data*, vol. 3, 2016, p. 160035, doi: 10.1038/sdata.2016.35.

[9] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, "Diagnosis code assignment: models and evaluation metrics," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 2, 2014, pp. 231–237, doi: 10.1136/amiajnl-2013-002159.

[10] M. Subotin, A.R. Davis, "A System for Predicting ICD-10-PCS Codes from Electronic Health Records," *Workshop on BioNLP (BioNLP)*, 2014, pp. 59–67.

[11] S. Abhyankar, D. Demner-Fushman, F. Callaghan, "Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 5, 2014, pp. 801–807.

[12] J. Pathak, K.R. Bailey, C.E. Beebe, S. Bethard, D.S. Carrell, P.J. Chen, et al, "Normalization and standardization of electronic health records for high-throughput phenotyping: The SHARPn consortium," *J. Am. Med. Inform. Assoc.*, vol. 20, no. e2, 2013, pp. e341–e348, doi: 10.1136/amiajnl-2013-001939.

[13] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, suppl. 1, 2004, pp. D267–D270, doi: 10.1093/nar/gkh061.

[14] RIZIV, *Rijksinstituut voor ziekte- en invalidite itsuitkeringen nomenclature*, http://www.riziv.fgov.be/NL/nomenclatuur/Paginas/default.aspx>.

[15] E. Scheurwegs, K. Luyckx, L. Luyten, W. Daelemans, T. Van den Bulcke, "Data integration of structured and unstructured sources for assigning clinical codes to patient stays," *J. Am. Med. Inform. Assoc.*, vol. 23, no. e1, 2016, pp. 11–19, doi: 10.1093/jamia/ocv115.

[16] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, no. April, 2016, p. 26094, doi: 10.1038/srep26094.

[17] R. Cohen, M. Elhadad, N. Elhadad, "Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies," *BMC Bioinform.*, vol. 14, no. 1, 2013, p. 10, doi: 10.1186/1471-2105-14-10.

[18] S.M. Vieira, L.F. Mendon□a, G.J. Farinha, J.M.C. Sousa, "Modified binary {PSO} for feature selection using {SVM} applied to mortality prediction of septic patients," *Appl. Soft Comput.*, vol. 13, no. 8, 2013, pp. 3494–3504.

[19] T. Botsis, M.D. Nguyen, E.J. Woo, M. Markatou, R. Ball, "Text mining for the Vaccine Adverse Event Reporting System: Medical text classification using informative feature selection," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, 2011, pp. 631–638.

[20] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, 2003, pp. 1157–1182, doi: 10.1016/j.aca.2011.07.027.arXiv:1111.6189v1.

[21] R. Kohavi, G.H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, 1997, pp. 273–324, doi: 10.1016/S0004-3702(97)00043-X.

[22] H.C. Peng, F. Long, C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, 2005, pp. 1226–1238.

[23] S. Fu, M.C. Desmarais, "Markov blanket based feature selection: a review of past decade," *Proceedings of the World Congress on Engineering 2010*, vol. I, 2010, pp. 321–328.

[24] I. Tsamardinos, L.E. Brown, C.F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, 2006, pp. 31–78, doi: 10.1007/s10994-006-6889-7.

[25] A.E.W. Johnson, T.J. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, et al, "MIMIC-III, A freely accessible critical care database," *Scientific Data*, 2016, doi: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635.

[26] B. Dolhansky, *Artificial Neural Networks: Linear Multiclass Classification (Part 3) September 27, 2013 in ml primers, neural networks*, http://www.briandolhansky.com/blog/2013/9/23/artificial-neural-nets-linear-multiclass-part-3.

[27] E. Scheurwegs, B. Cule, K. Luyckx, L. Luyten, W. Daelemans, "Selecting relevant features from the electronic health record for clinical code prediction," *Journal of Biomedical Informatics*, vol. 74, 2017, pp. 92–103.

[28] S.M. Zhou, F. Fernandez-Gutierrez, J. Kennedy, R. Cooksey, M. Atkinson, S. Denaxas, C. Sudlow, "Defining disease phenotypes in primary care electronic health records by a machine learning

approach: a case study in identifying rheumatoid arthritis," *PloS one*, vol. 11, no. 5, 2016, p. e0154515.

[29] T. Lingren, P. Chen, J. Bochenek, F. Doshi-Velez, P. Manning-Courtney, J. Bickel, et al, "Electronic health record based algorithm to identify patients with autism spectrum disorder," *PloS one*, vol. 11, no. 7, 2016, p. e0159621.

[30] J. Zhao, P. Papapetrou, L. Asker, H. Boström, "Learning from heterogeneous temporal data in electronic health records," *Journal of Biomedical Informatics*, vol. 65, pp. 105–119.

[31] S. Pai, S. Hui, R. Isserlin, M.A. Shah, H. Kaka, G.D. Bader, "netDx: interpretable patient classification using integrated patient similarity networks," *Molecular Systems Biology*, vol. 15, no. 3, 2016, p. e8497.

[32] P. Courtiol, C. Maussion, M. Moarii, E. Pronier, S. Pilcer, M. Sefta, T. Clozel, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Medicine*, vol. 25, no. 10, 2019, pp. 1519–1525.

[33] B. Dolhansky, 2013, *Artificial neural networks: Mathematics of backpropagation (part 4)*, https://www.briandolhansky.com/blog/2013/9/27/artificial-neural-networks-backpropagation-part-4.

[34] B. Dolhansky, J.A. Bilmes, "Deep submodular functions: Definitions and learning. Advances in Neural Information Processing Systems," vol. 29, 2016, pp. 3404–3412.

[35] G. Paliwal, A. Bunglowala, P. Kanthed, "An architectural design study of electronic healthcare record systems with associated context parameters on MIMIC III," *Health and Technology*, 2022, pp. 1–15.

[36] H. Sharma, C. Mao, Y. Zhang, H. Vatani, L. Yao, Y. Zhong, Y. Luo, "Developing a portable natural language processing based phenotyping system," *BMC Medical Informatics and Decision Making*, vol. 19, no. 3, pp. 79–87.

[37] A.P. Reimer, A. Milinovich, "Using UMLS for electronic health data standardization and database design," *Journal of the American Medical Informatics Association*, vol. 27, no. 10, 2020, pp. 1520–1528.

[38] H. Zhang, T. Lyu, P. Yin, S. Bost, X. He, Y. Guo, J. Bian, "A scoping review of semantic integration of health data and information," *International Journal of Medical Informatics*, 2022, p. 104834.

[39] D. Yuvaraj, A.M.U. Ahamed, M. Sivaram, "A study on the role of natural language processing in the healthcare sector," *Materials Today: Proceedings*, 2021.

[40] H. Sharma, C. Mao, Y. Zhang, H. Vatani, L. Yao, Y. Zhong, et al, "Developing a portable natural language processing based phenotyping system," *BMC Medical Informatics and Decision Making*, vol. 19, no. 3, 2019, pp. 79–87.

[41] S. Moosavinasab, E. Sezgin, H. Sun, J. Hoffman, Y. Huang, S. Lin, "DeepSuggest: Using neural networks to suggest related keywords for a comprehensive search of clinical notes," *ACI open*, vol. 5, no. 01, 2021, pp. e1–e12.

[42] B. Wang, Y. Sun, Y. Chu, D. Zhao, Z. Yang, J. Wang, "Refining electronic medical records representation in manifold subspace," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–17.

[43] E.O. Omuya, G.O. Okeyo, M.W. Kimwele, "Feature selection for classification using principal component analysis and information gain," *Expert Systems with Applications*, vol. 174, 2021, p. 114765.

[44] U.A. Bhatti, L. Yuan, Z. Yu, S.A. Nawaz, A. Mehmood, M.A. Bhatti, et al, "Predictive Data Modeling Using sp-kNN for Risk Factor Evaluation in Urban Demographical Healthcare Data," *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 1, 2021, pp. 7–14.

[45] N. Wang, Y. Huang, H. Liu, Z. Zhang, L. Wei, X. Fei, H. Chen, "Study on the semi-supervised learning-based patient similarity from heterogeneous electronic medical records," *BMC medical informatics and decision making*, vol. 21, no. 2, 2021, pp. 1–13.

[46] C. Comito, D. Falcone, A. Forestiero, Diagnosis prediction based on similarity of patients physiological parameters. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 487–494.

[47] W.M. Shaban, A. Rabie, A.I. Saleh, M.A. Abo-Elsoud, "Accurate detection of COVID-19 patients based on distance biased Na□ve Bayes (DBNB) classification strategy," *Pattern Recognition*, vol. *119*, 2021, p. 108110.

[48] V. Jackins, S. Vimal, M. Kaliappan, M.Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing*, *77*(5), 2021, pp. 5198–5219.

[49] J.M. Bae, "The clinical decision analysis using decision tree," *Epidemiology and health*, vol. *36*, 2021.

[50] S.H. Rukmawan, F.R. Aszhari, Z. Rustam, J. Pandelaki, "Cerebral infarction classification using the k-nearest neighbor and naive bayes classifier," *Journal of Physics: Conference Series*, Vol. 1752, No. 1, 2021, p. 012045.

[51] L.I. Qi, "Patient classification with ensemble treebased modelling for decision support in acute clinical care settings," Doctoral dissertation, RMIT University.

[52] A. Singh, A. Dhillon, N. Kumar, M. Hossain, G. Muhammad, M. Kumar, "eDiaPredict: An

Ensemble-based framework for diabetes prediction," *ACM Transactions on Multimidia Computing Communications and Applications*, vol. 17, no. 2s, 2021, pp. 1–26.

[53] T. Razzaghi, O. Roderick, I. Safro, N. Marko, "Multilevel weighted support vector machine for classification on healthcare data with missing values," *PloS one*, vol. 11, no. 5, 2021, p. e0155119.

[54] D.M. Abdullah, A.M. Abdulazeez, "Machine Learning Applications based on SVM Classification A Review," *Qubahan Academic Journal*, vol. 1, no. 2, 2021, pp. 81–90.

[55] N. Shahid, T. Rappon, W. Berta, "Applications of artificial neural networks in health care organizational decision-making: A scoping review," *PloS one*, vol. 14, no. 2, 2019, p. e0212356.

[56] W. Liu, Z. Wang, N. Zeng, F.E. Alsaadi, X. Liu, "A PSO-based deep learning approach to classifying patients from emergency departments," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 7, 2021, pp. 1939-1948.

[57] B.S. Panchbhai, V.M. Pathak, "A Systematic Review of Natural Language Processing in Healthcare," *Journal of Algebraic Statistics*, vol. 13, no. 1, 2022, pp. 682–707.

[58] B. Zhou, G. Yang, Z. Shi, S. Ma, "Natural language processing for smart healthcare," *IEEE Reviews in Biomedical Engineering*, 2022.

[59] K.M. Al-Aidaroos, A.A. Bakar, Z. Othman, "Medical data classification with Naive Bayes approach," *Information Technology Journal*, vol. 11, no. 9, p. 1166.

[60] I. Korobiichuk, A. Ladanyuk, R. Boiko, S. Hrybkov, "Features of Control Processes in Organizational-Technical (Technological) Systems of Continuous Type," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 14, no. 4, pp. 11–17. doi: 10.14313/JAMRIS/4-2020/39.

[61] S. Yousfi, M. Rhanoui, M. Mikram, "Comparative Study of CNN and LSTM for Opinion Mining in Long Text," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 14, no. 3, pp. 50–55. doi: 10.14313/JAMRIS/3-2020/34.

[62] A. Ndayikengurukiye, A. Ez-zahout, A. Aboubakr, Y. Charkaoui, O. Fouzia, "Resource Optimisation in Cloud Computing: Comparative Study of Algorithms Applied to Recommendations in a Big Data Analysis Architecture," *Journal of Automation, Mobile Robotics and Intelligent Systems*, vol. 15, no. 4, pp. 65–75. doi: 10.14313/JAMRIS/4-2021/28.