**Use of Deep Learning to Extract Health Informatics From HER**

**Authors:**

*Ms. Mayuri Diwakar Kulkarni*

Assistant Professor

Dept of Computer Engineering,

SVKM’s Institute of Technology, Dhule

*Mr. Ashish Suresh Awate*

Assistant Professor

Dept of Computer Engineering,

SVKM’s Institute of Technology, Dhule

*Dr Makarand Shahade*

Associate Professor

Dept of Computer Engineering,

SVKM’s Institute of Technology, Dhule

***Abstract:***

***Information plays a very crucial role in the Healthcare sector. To know patient health condition every in details of patients is required. This health information is sensitive as health is concerned. Hence preserving that information is very important. This information is stored in a form of an electronic record in servers, flash memories etc and known as Electronic Health Record.***

***This EHR stores clinical reports such as prescriptions, pathology tests, x-rays, MRIs, etc. This information is very huge and also unstructured. Hence retrieving this information is very essential for decision making. Making decisions from EHR is very complex due to huge and heterogeneity and unstructured manner of organization of data. Here comes the role of deep learning. For this deep learning provides a solution based upon present features in it such as processing on heavy workload for a huge amount of data using GPU along with a Layered model for classification, overfitting, and underfitting issues handled using hyper-parameters, minimized cost function.***

***These features help in decision-making from the available EHRs. These EHRs will help healthcare workers to predict the risk of associated disease. These prediction models are built using various DL models such as FFNN, CNN, RNN, etc.***

***The chapter contains the role of EHR in the healthcare sector as well as how DL models are helpful to extract health informatics from the EHR.***

***Keywords: Deep Learning, EHR, FFNN, RNN, CNN, Classification, unstructured data, Health informatics***

 **Introduction:**

As we all are aware that Deep learning is the subset in the main stream diverse technology as Artificial Intelligence. Deep learning doesn't require preprocessing on the data set because of its architectures layer. This architecture layer is responsible to analyze data. But how does this happen in Deep learning?

Deep learning models consist of layered structure. In which the first and prime concerned layer as input layer. This layer provides the input to the next layer as hidden layers. These layers are responsible to perform the processing and provides the outcome to the next layer called output layer. These layers in deep learning are used to do analyze data.

As the main three layers in deep learning, the input and output layer are single-layered separately whereas the hidden layer is hidden as well as it uses more than one layer in it. Hence models in deep learning use cascading of multiple layers. These hidden layers are responsible to filter data. These layers are cascaded with multiple layers with each successive layer taking input from the previous layers. Deep learning has input layered fed to hidden layers. The hidden layer is fed to the output layer.

The input layer takes input features from the dataset and is applied to the hidden layer. This hidden layer trains through hidden neurons which are supplied to the output layer. Output layer responsible for the classification. The presence of the number of hidden layers in the model is directly proportional to the complexity of the problem.

The main motto of utilizing these deep learning models are to deal with the huge amount of data along with the diversity in the data types. Hence it is generally seen in deep learning that uses heterogeneous data. This data is text, images, and videos. The fundamental difference between machine learning and deep learning is that machine learning works on usually text data or structured data and if it uses images or videos need preprocessing before it is given to the training model whereas deep learning does not require preprocessing because of the hidden layer.

The learning algorithms in Deep learning uses the labelled data as a supervised learning for classification problems, unsupervised, semi-supervised learning. Supervised learning uses categorical data which is labeled, unsupervised learning uses categorical data which is not labeled whereas semi-supervised learning algorithm uses categorical which is labeled as well unlabeled[48].

Neural network architecture is adopted by Deep Learning. The common Supervised learning algorithms in deep learning algorithms are Artificial Neural Network (ANN), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN).

ANN is a component of computing systems such analyzes the human brain and makes a decision. In ANN, neurons are responsible for computing information that is supplied to hidden neurons and given to the output layer. ANN identifies a pattern in the training phase at initial stage as the input layer, the output after the input layer which is trained is compared with the actual output, the difference between these outputs produces an error. This algorithm uses tabular data, image, and Text data. But the limitations of ANN are to work only on a 1-D vector for training purposes to model. Due to this 1-Dimensional vector data, the drawback as increased parameters increases the size of images. This algorithm cannot capture input data as sequential information.[1]

RNN algorithm is used to extract information from the time series data, audio data, or text data. This captures sequential input as the data and predicts the dependency present in a sequence. This shares the different times using parameter sharing. However, RNN suffers during backpropagation because of the gradient at the input layer which is obtained by multiplying the gradient of successive layers.[2]

CNN algorithm is used to process image and video processing. This automatically filters the data to extract information from the inputs. It extracts spatial features from the image. This is used to extract the object and location of the object.[3]

CNN uses multiple layers to process inputs and extract essential features from the image using activation function as ReLu, Pooling to reduce dimensions, Flattening uses pooling output into the one-dimensional matrix which is given to a fully connected layer.

RNN takes input from the previous layer and is supplied to the current step. This is suited for sequential data. RNN has four steps.[4]



Fig 1. Characteristics of DL

The diagram shows the important characteristics of deep learning. These characteristics in deep learning make it more prominent to use it on an unstructured and huge amount of data by using a good learning algorithm along with optimizing hyper-parameters which also needs an error-free system by considering effective cost functions.

Deep learning is mainly used in various sectors where the data is collected from heterogeneous data sources such as sensors, cameras or various web sources, social media, etc. This data is continuously generated hence it is so huge in capacity. Only collecting the data is not at all essential. To extract the information from that is very important. This extracted data is very important to predict, forecast, or in recommendation systems.

Based on this application area, deep learning plays a very important role to extract information from Electronic Health records. This extracted information is essential to healthcare works either for diagnosis or prediction. This information is also useful in care centers as well as used by the drugs suppliers to know the effects and side effects of the medicine.

Some questions such as what is Information Extraction, what is EHR, and how deep learning is used to extract information from EHR are very important for the discussion. Hence the paper focuses on Information Extraction, EHR, data types in EHR, and finally deep learning to extract information.

**Information Extraction:**

Information Extraction is a process through which essential information is extracted or retrieved and converted to structured data formats. There are various techniques to extract information from the unstructured data are tokenization, PoS tagging, Dependency Graph, NER with Spacy.

Tokenization separates text into smaller units called tokens. This tokenization is categorized into three types of words, characters, and sub-words called n-gram characters. SOTA, RNN, GRU, LSTM were used to preprocess the data at token levels in a transformer-based model. Word tokenizer separates text into individual words using word2vec, GloVe called as word embedding in tokenization. This method suffers from Out of Vocabulary (OOV). In character tokenization words are split into sets of characters this overcomes drawbacks in word tokenized of OOV but it increases the length of input as well as output which represents the sentence as the character which has a proper sequence. The text is separated into sub-words called sub-word tokenization using SOTA and BPE[45].

POS is a supervised learning approach that classifieds the words into parts of speech called lexical categories using MLP. This contains input data that is preprocessed which later extracts features and is applied to the classifiers as CNN, Bi-LSTM, Hybrid models, and results analyzed based on evaluation matrices[47].

**What is EHR:**

***EHR:***

EHR is electronic health records that stores patients report which are clinical reports. These reports are MRI, X-ray, Pathology reports, CT-scan, etc, or prescriptions prescribed by the doctors. These records are traditionally stored on paper but due to technological advancements, it is started to store data on electronic platforms since 1960. But this terminology came into the picture due to its availability and affordability in 1980.[5]

In the late 1980 and early 1990 due to easy accessibility of the hardware such as computers and Internet medical records stored on local disk or web-based EHR. This data was stored on the papers but due to the modern approach people started to preserves data on electronic media. These electronic media are CD, DVD, pen drives, Cloud storage or NFC, etc. This electronic form of data is accessible anywhere in the world. This EHR is useful in clinical decision support[6].

***Types of EHR:***

The following figure depicts the types of EHR

* Physician-Hosted System :

In this type of system, the software and hardware required to implement an EHR system are owned by the physician himself. It is useful for a physician who has a very large scale of medical practice[30].

**Advantage :**

* It is secure as the record is stored in a physician system hence without authorization it is not possible to retrieve any information from the system.
* Reliable.

**Disadvantage :**

* Complex to handle large software.
* Costly: As the physician is only responsible to make the store, maintaining, and updating the system, needs to pay for the services to the facilitators. This software designed is customized and implemented on the local servers.



**Fig 2.Types of EHR[32]**

* Remotely-Hosted System :

In this type of EHR, the storage of data and its maintenance is handled by a third party. This is useful and attractive for physicians having limited practice[30].

Advantage :

* Cheap compared to the physical EHR system
* Less overhead of hardware purchase and maintenance

Disadvantage :

* The initial cost is high
* Increase dependency for data on the remote system

 3. Remote System :

Remote System allows physicians or partitioners to access data remotely either in various forms such as subsidized, dedicated, and cloud. These remote systems types are categorized as follows:

1. **Subsidize:** In this, the physician or practitioner ties up with big hospitals and allows them to store and maintain the data for the EHR system which in turn subsidizes the cost of EHR for them.
2. **Dedicated:** In a dedicated remote system, physicians host their EHR data on a dedicated vendor’s server and most of the data-related work is handled by that vendor.
3. **Cloud:** In a Cloud-based remote system rather dedicated server the data is stored on an internet-based distributed system. This is the most popular system amongst the EHR systems. In this system, data is always available from anywhere through the vendor's website.[30]

Choosing the best EHR system is based on the requirement of the physician but nowadays the high-speed internet availability makes the cloud-based system popular among the physician and hospitals which can’t afford the infrastructural cost of a premises EHR system. But still, the large healthcare systems rely on a physical hosted system as it is more secure and gives them full control over data.

**Advantages of EHR:**

As EHR is an electronic way to carry health information. This electronic means of the information-carrying system provides the following advantages such as

* An accurate record is required at the time of patient care. This accurately carries very critical information.
* Due to the availability of data on the internet platform, it is easy to access data from any point. This helps to quick access of data from any place from any location. This results in easy coordination of the data.
* This reduces error which results in the inaccurate diagnosis of the disease results in proper treatment.
* Due to proper data maintenance, the data is stored in an appropriate format which helps to communicate details. These details are properly maintained in EHR which results in communication convenience to healthcare practitioners.



**Fig3 Advantage of EHR**

**Modules in EHR:**

To have data on the electronic form is not only the prominent solution unless the details for the data are not present in systems. The associated modules generated data for EHR areas Administrative data, Nursing Model, Laboratory Data Model, Radiology Images, Clinical Documents, Pharmacy. These modules are explained below in detail as:

* **Administrative Data:**
1. Patient Details :

These patients' details are blood group, weight, a height which are required for the nursing purpose. But in some of the circumstances demographics, race, eating habits, allergies, etc are required. Hence from a healthcare point of view for nursing and treatments, these details are required to practitioners.

1. Hospital Admission Details:

In this, the details of admission to hospitals and care have taken record is maintained. This also stores the administrative data. This data is required for hospital authorities to keep track of patients either financial or patient care information.

* **Nursing Model:**

The nursing model of EHR contains electronic data related to the patient such as the medical history of the patient, allergies, plan of treatment, medication history, diagnosis, the date for immunization, test results, etc.[27][28]. This provides a convenient way to decide a plan for patient care while nursing. In other words, this model keeps all clinical electronic records required for patient care during his/her admission to the hospital.

* **Laboratory Data Model:**

Laboratory data model stores Clinical data such as microbiology and biochemistry. This provides the initial data to medical partitioners for examination purposes.

* **Radiology Images**:

This data is in a form of images. These images are x-ray, CT scan, and MRI. This data helps practitioners to examine patients for long-term and critical diseases.

This data is generated through the machinery. Hence it is recommended by the researchers to predict the prevalence of critical circumstances of any disease or to diagnose the associated difficulty with patients to provide machinery details such as precision and accuracy parameters along with the tolerance levels, encoding schemes, etc.

Based upon these recommendations, reports or images specify all these terminologies related to machinery embedded in the report. Due to this, the burden of repetitive tests on patients reduces. This also helps patients to reduce the financial burden during treatment.

* **Clinical Documents:**

Clinical documents are the documents such as prescriptions, medical notes data, treatment plans. This also contains details of daily updates of patients based upon the treatment of course.

* **Pharmacy:**

This Pharmacy data stores the medication history as drug code, drug name, and course for the medication. These details are essential to decide the way of medication for the patients. This also decides the treatment flow. This also stores the information of the allergies and reactions due to drugs and their course.

This helps pharmacists to help in the research area for drug discovery as well drug development.

The figure shows various components that provide the data to the healthcare practitioners. Hence it is also called a patient's electronic health record. This record consist of various data such as text, images, videos, etc. This record is very important in critical circumstances to the doctors.

 

**Fig 4 Components of EHR**

The data EHR required in the following circumstances:

* Information extraction: To retrieve required data from stored in EHR.
* Disease Diagnosis: To diagnose disease from the patient's EHR.
* Disease prediction: To predict disease based upon available data from the clinical, radiological data.

Now, this store records in the systems and collected them through various data collection sources and later stored them in a system at cloud. This stored data in the cloud is very huge as well as heterogeneous in nature. Hence retrieving information from the huge amount of data is very essential. For this extraction of information deep learning plays a very significant role. The information extraction methods are categorized as rule-based methods, classification-based methods, sequential labeling Based methods, and Hybrid models[46].

These extraction methods are explained below in detail:

* 1. **Rule-based Method:**

Rule-Based information extraction systems contain rules and interpreters to interpret rules. These rules are patterns to be fulfilled by the position in documents. These rules are specified in regular expressions. These regular expressions use the sequence of characters that finds search patterns. This is used when data is in a form of text. The text data is available in clinical reports as well as in clinical documents. These data stores reports, prescriptions, line of course of medicine, and results. Hence these rules are to be developed in two ways. These ways are either manual knowledge and leveraging knowledge-based systems or hybrid systems.

This rule-based method extracts information from a semi-structured web page. It learns rules as syntactic or semantic barriers in text. This uses the top-down method and bottom-up approach[7].

* 1. **Classification-based Method:**

The classification-based Method is a supervised learning approach. This approach uses labeled data used to extract information. This information extraction using the classification method uses two phases. These two phases to extract information from data are as follows:

* Learning: This uses labeled data from pre-existing documents to predict the future from the learning model.
* Extraction: This phase accepts learned models to label from unlabeled documents.

Based upon these two-phase models, classification of the disease can easily be possible from the huge data. Deep learning models are used to classify diseases from huge datasets as well as the heterogeneity of the data in EHR. That heterogeneity of data collects clinical reports from the various groups. This uses different learning classification algorithms. These algorithms are CNN, LSTMs, RNN, GAN, RBFN, MLP, SOM, DBN.

* 1. **Sequential Labelling Based Methods:**

Sequential labeling is an NLP task. This assigns labels to each token. These tokens are considered for searching purposes. Token labeling and span labeling are the major types of sequential labeling. The difference between token labeling and span labeling is that token labeling uses individual POS whereas span labeling uses a group of words known as tags[44].

Based upon these two types of sequential labeling based methods to retrieve the information. This retrieval of information from the stored records such as clinical reports, prescriptions, medical records, and patient information is raw data. This uses information extraction principles as :

* Dependencies in the source information.
* Differentiating factor to observe proper dependency data
* Accuracy measures the information from the data.

 HMM and CRF is the best models for sequential labeling-based methods that take sequential input instances and predict an optimal sequence of labels. But these approaches suffer due to a lack of semantic[26].

 **4. Hybrid Models:**

These models use a combination of more than one approach to extracting information from the data. The semantics involves two components in information extraction using information models and value sets. The information model specifies data semantics and defines the relationship, rules, and operations. The value sets to store the permissible values.

**Data Types in EHR**

EHR data types are categorized as common data types and emerging data types. Common data types are

1. **Demographics Data: This** demographic data includes general but very essential terms such as age, gender, and medical parameters are vary concerning age and gender along with race.
2. **Diagnoses** **Data**: This data includes the medical history of the patients as well as diagnosed disease severity.
3. **Problem** **List Data:** This data contains the status of diagnosed disease as an active or resolved diagnosis.
4. **Data of Family History:** This data includes hereditary medical issues and their severity.
5. **Allergies Data:** The Allergies data includes medicinal allergies and food allergies.
6. **Immunization Data:** This data contains DTaP, HepB, IPV
7. **Medication Data:** In this data type it includes written prescriptions by the experts, a record of doses its frequency, and its course of duration.
8. **Procedures Data:** This data type includes inpatient and outpatients. American Medical Association standardizes services in healthcare sectors using current procedural terminology called CPT. These services are in medical, surgical, or diagnostic aspects. The CPT codes store three types of procedural data as
* **CPT Category 1 codes** are codes for Evaluation & Management, Anesthesia, Surgery, Radiology, Pathology and Laboratory, Medicine services, and procedures.
* **CPT Category 2 codes** are for Patient management, patient history, physical examination, Diagnostic or screening Processes, Follow-up, Therapeutic, preventive, etc.
* **CPT Category 3 codes** for new technologies, services, and procedures[8].

Hence the CPT code sets are HCPCS level II for procedures, services, drug supplies, and equipment details required during medical diagnosis or examination.ICD-10 PCS specifies facilities inpatient procedures. ICD 10-CM diagnosis of patients of inpatient or outpatients provides.

1. **Lab orders Data:** This data contains laboratory reports as Blood Test reports.
2. **Vital Sign Data: Vital sign data** consist of Body Mass Index as weight, height, and Blood Pressure.
3. **Data type as a Report: Reports containing** radiology, pathology, etc.
4. **Utilization Data:** This includes data of admission to the hospital and treatment costs.

**Emerging Data Types:**

Some emerging data types are Biosample Data, Genetic Information, Social data, Patient-Generated Data, Community Data, Geo-Spatial, Survey, Free Text [34].

EHR data is useful in the following circumstances :

* **Biosample Data:** This data consists of race, age, gender of the patient
* **Genetic Information:** This maintains the record of medical history, side effects, and diagnosis.
* **Social Data:** This data maintained the main two types of diagnosis record of patient as active and resolved diagnosis
* **Patient-Generated:** This data maintains the known disorders in the patient and the associated risk factors
* **Community**: It contains data related to allergies caused by food or medication and anaphylaxis
* **Geo-spatial:** it contains geographical indications such as DTaP, HepB, IPV
* **Surveys**: It contains the written Prescriptions.
* **Free Text:**  inpatient, outpatient
* **Other Data Types:** it contains mainly general laboratory test data such as CBC test results, level of HbA1C.

**Role of Deep Learning to extract information:**

Deep learning approach to extract information from various data types. These data types are demographic databases that store the actual maps and locations. The other forms of data collection from IoT devices as real-time inputs. Hence it is huge in storage. To extract information from various data modules discussed below.

The linguistic model extracts information using an unsupervised learning approach. This extracts information for a very specific and small data set in the training phase. This Linguistic Model uses ULMfit, ELMo. These are unsupervised algorithms that embed large text corpora to gain knowledge using RNN[9].

ULMfit is Universal Language Model Fine-Tuning for Text Classification. This uses an inductive learning approach for task-based modification. This classifier uses the activation function as ReLu in the hidden layer and softmax in the output layer. This classification technique reduced error by 18-20%[10].

ELMo ("Embeddings from Language Model") uses context from the words. This represents a sequence of words using the word embedding method as a corresponding sequence of vectors. It used Character-level tokens and fed them to bi-directional LSTM as an input, which in results produces word-level embeddings.

The technique in NLP, Named Entity Recognition (NER) is used to extract information from the descriptive information. This descriptive information in the EHR system in administrative System modules such as Patient registration details and Hospital Records as well as in Nursing Module such as Patients Details. This NER technique uses supervised, unsupervised, and semi-supervised algorithms. This technique uses semantics.

This is used to extract information from the clinical reports. This technique determines the quantity and type of drugs for the treatment[11][12][13].

**Information Extraction approaches for different Data Sets:**

Various modules in EHR systems stores data in various forms. As per the requirement of the data, module data is stored. This data is required for various purposes. The approaches used to extract information from various data in EHR are explained below in detail:

***Demographic Data Type:***

To examine the health annually to detect tuberculosis in chest radiograph using demographic information. This demographic variable such as age, gender, height, weight used as input along with the x-rays. These demographic variables are inputted to the Hidden Layer in CNN to diagnose the probability of tuberculosis. These demographic features were extracted using Global Average Pooling (GAP). Later SGD optimized was used. AUC differences compared with the CNN model only for radiology data with demographic variables. By inclusion of demographic variables performance of the system results in more promising concerning AUC and sensitivity of the model also increases[14].

**Medication Data**

The Chinese medical terms are evaluated using a statistical model which finds the relationship between the events such as the admission, course of treatment, course of medication, discharge information, allergies, etc are in text form, this needs to be driven by recognition using NER. The bag-of-character feature for conditional random field proposed by the Chinese EMR and validation is done using 5-fold cross-validation to evaluate the CRF method. To match the description in EHR data is classified using binary support vector machine. A medical event and description use local context features uses three words before/after and recognized terms/candidate description, as well as semantics features, describes the normalized google distance between medical term and candidate description[15].

Multitask RNN with contextual word embedding model BERT improvises the NER tasks. These NER tasks are divided into two parts as named entity discovery and named entity classification[16].

**Information Extraction from EHR using Deep Learning**

Deep Learning is used to discover and stratify new subtypes and improve classification under existing diseases. This uses data which is taken from an EHR module that stores the clinical records. This method evaluates information and shares quantitative and qualitative components. The medical discharge summaries, diagnostic test reports are written in textual format This EHR uses Name Entity Recognition and Relation Extraction[17].

These two extraction methods. To extract information from the written textual format requires NLP tools.

NER system uses a deep learning model. In this model, the input is the raw data as a sequence of the words in the sentences. The output from the input layer is given to the embedded layer. This layer transforms the sequence of the word into density-valued vectors. Deep Neural networks train wide-level feature representation and feature long-term dependencies.[18][more things to be included].

NER system uses CNN architecture uses Hyperbolic Tangent Action Function (HardTanh) as non-linear layers and standard linear layers. As the input, the system is text data hence to the English layer NLP tasks are applied. This captures words using an embedded layer and later mapped to an N-dimension vector. Long-term dependencies in hidden layers are captured using a convolution layer. Stochastic gradient descending is used to train Neural networks for local and global features. Finally, these extracted features for supplied to the classification layer to classify the sentence level log-likelihood approach[19].

In the proposed work[20] hybrid systems are used to extract information related to the drug. This extracts data from the dosage, frequency, and duration of the medication along with route condition. Skip-gram, FastText embedding, and embedding for language model tested on Embedding layer with bidirectional LSTM and bidirectional LSTM conditional random field.

CNN uses a Bag of words, n-gram, and embedding-based logistic regression model used for the feature representation[21].

NER was used to identify the predefined entities from text to extract clinical information at the initial stage later this initial extracted information using different models such as BiLSTM-CRF, BERT, and BERT-CRF. CT reports are trained using BiLSTM-CRF.For optimization purposes Hyper-parameter for BiLSTM-CRF Mini -batch stochastic gradient descent with learning rate 0.1 to 0.9[22].

CNN architecture is designed for specific tasks in computer vision and medical imaging. This model trains the 2D,3D model of CT or MRI by cross-sectional abdomen image[23].

NER and Relational extraction are the two stages in deep learning that are used to extract text from the medical data. NER uses Bi-LSTM and CRF ,for relation extraction CNN model is applied[27].

**BioSample Data:**

BioSample data contain data about a single biospecimen that has attributes such as age, cell types, gender, genome types, tissues strain, etc. Hence it uses the NLP approach for NN to convert free text to numerical data. This conversion from text to numeric using the word2vec model. This uses bi-LSTM for training as well as testing data sets with 64 hidden layers[24].

To extract information from the biosample data used technique as word embedding for information retrieval using vectorization using a one-hot encoding scheme. Similarities in words are estimated using semantic evaluation. This classifies short phrase entities[25].

**Genomic Data:**

Genomic data is about an individual DNA. This data consists of the genetic test as well as biospecimens. Genomic data used to link clinical data for diagnosing diseases. This connection useful in predicting associated risk for disease along with way of patients response based upon treatment and the reactions of the drugs. The approach used by researchers to extract features from the genomic dataset is to have a DNA strands model for the scanning and performance measure used within the framework[33].

Various Deep Learning approaches for EHR

* **eNRBM**: In this approach, the author proposed the model based on Auto Encoder Architecture as eNRBM (“EMR-driven nonnegative restricted Boltzmann machines”)[35] to predict suicide risk. They use a modular learning process for the 7,578 patient records which consist of attributes such as
	1. Diagnoses - International Statistical Classification of Diseases and Related Health Problems-10
	2. Procedures - Australian Classification of Health Interventions
	3. Elixhauser comorbidities
	4. Diagnosis-related groups
	5. Emergency attendances and admissions,
	6. demographic variables are ages in 10-year intervals and Gender

	This model is tested on performance metrics such as F1-score, precision, and recall
* **Deep Patient:** In this approach, the author proposed the Deep patient [36] model which uses stacked denoising autoencoder architecture for predicting future diseases. They use a modular learning process for the 704,587 patient records which consist of attributes such as
	1. Demographic variables are age, gender, race,
	2. diagnoses - International Statistical Classification of Diseases and Related Health Problems-9
	3. Medications data
	4. Procedures data
	5. Lab tests
	6. free-text clinical notes

This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC), Accuracy, F1-score, Precision.

* **Deepr:** n this approach the author proposed the Deepr model which uses Convolutional Neural Network (CNN) architecture for predicting unplanned readmission. They use an end-to-end learning process for the 300,000 patient records which consist of attributes such as
1. Diagnoses - Australian Coding Standard
2. procedures - Australian Classification of Health Interventions

This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC).

* **DeepCare:** In this approach, the author proposed the DeepCare[38] model which uses Recurrent Neural Network (RNN) architecture for predicting unplanned readmission and disease progression. They use an end-to-end learning process for the 7,191 patient records which consist of attributes such as
	1. Diagnoses - International Statistical Classification of Diseases and Related Health Problems (ICD-10)
	2. procedures - Australian Classification of Health Interventions
	3. medications - Anatomical Therapeutic Chemical Classification
* This model is tested on performance metrics such as F1-score, precision, and recall

 **Med2Vec:** In this approach, the author proposed the Med2Vec [39] model which uses Feed Forward Neural Network (FFNN) architecture for predicting medical codes in previous/future visits. They use an end-to-end learning process for the not disclosed number of patient records which consist of attributes such as
	1. Diagnoses - International Statistical Classification of Diseases and Related Health Problems (ICD-9),
	2. procedures - Current Procedural Terminology
	3. medications - National Drug Codes

	This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC) and recall
* **Doctor AI:** In this approach, the author proposed the Doctor AI [40] model which uses Recurrent Neural Network (RNN) architecture for predicting medical codes in future visits or duration until the next visit. They use an end-to-end learning process for 263,706 patient records which consists of attributes such as
	+ Diagnoses - International Statistical Classification of Diseases and Related Health Problems (ICD-9)
	+ procedures -Current Procedural Terminology
	+ medications - Generic Product Identifier

	This model is tested on performance metrics - R2 score and recall
* **RETAIN:** In this approach, the author proposed the RETAIN[40] model which uses Recurrent Neural Network (RNN) architecture for predicting medical codes in future visits or duration until the next visit. They use an end-to-end learning process for 32,787 patient records which consists of attributes such as
	+ Diagnoses - International Statistical Classification of Diseases and Related Health Problems (ICD-9)
	+ procedures -Current Procedural Terminology
	+ medications - Generic Product Identifier

	This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC).
* **RETAIN:** In this approach, the author proposed the RETAIN[41] model which uses Recurrent Neural Network (RNN) architecture for predicting medical codes in future visits or duration until the next visit. They use an end-to-end learning process for 32,787 patient records which consists of attributes such as
	+ Diagnoses - International Statistical Classification of Diseases and Related Health Problems (ICD-9)
	+ procedures -Current Procedural Terminology
	+ medications - Generic Product Identifier

	This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC).
* **retains:** In this approach, the author proposed the RetainVis [42] model which uses Recurrent Neural Network (RNN) architecture for predicting cataract and heart failure. They use an end-to-end learning process for 63,030 for heart failure patient records and 117,612 for cataract patient records which consist of attributes such as
	+ Diagnoses - Korean Statistical Classification of Diseases and Related Health Problems
	+ procedures
	+ medications

	This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC)
* **Ensemble model:** In this approach, the author proposed the Ensemble model [43] model which uses ensemble architecture consisting of a weak learner of Feed Forward Neural Network (FFNN) and Recurrent Neural Network (RNN) for predicting inpatient mortality, 30-day unplanned readmission, long length of stay, diagnoses. They use an end-to-end learning process for 216,221 patient records which consists of attributes such as
	+ Demographics
	+ provider orders
	+ Diagnoses
	+ Procedures
	+ Medications
	+ laboratory values
	+ vital signs
	+ flowsheet data
	+ free-text medical notes

	This model is tested on performance metrics - Area Under the Receiver Operating Characteristic Curve (AUROC).

	**Conclusion:**

As EHR generates huge and heterogeneous data. This data is generated from various modules present in EHR.EHR is maintained in a remote place, remotely hosted storage, or cloud. These storage systems have their advantages associated with them. So the utilization of their types decides based upon applicability. The data in EHR are text, demographic data, BioSamples, Diagnostic data, clinical data, or radiological data. This data is in text, images, or videos. So the data in EHR must be stored in standard data patterns. If that standard storage pattern of data is maintained by all practitioners then Deep learning results in the best extraction of the information. So to extract information various deep learning methods are discussed and these methods are applied to the various types of data to extract information. This information extraction method uses neural networks and information processing is done by the NLP. This approach helps in information extraction from various modules present in an EHR system. Also, different learning models such as CNN, RNN, DNN are used to Predict the disease. Also, the chapter discusses various deep learning models to predict diseases.

**References:**

[1] Analytics Vidhya. 2022. *ANN vs CNN vs RNN | Types of Neural Networks*. [online] Available at: <https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/> [Accessed 30 January 2022].

[2] Analytics Vidhya. 2022. ANN vs CNN vs RNN | Types of Neural Networks. [online] Available at: <https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/> [Accessed 30 January 2022].

[3] Analytics Vidhya. 2022. *ANN vs CNN vs RNN | Types of Neural Networks*. [online] Available at: <https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/> [Accessed 30 January 2022].

[4] Analytics Vidhya. 2022. *Supervised Deep Learning Algorithms : Types and Applications*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/05/introduction-to-supervised-deep-learning-algorithms/> [Accessed 30 January 2022].

[5] Net Health. 2022. *The History of Electronic Health Records (EHRs)*. [online] Available at: <https://www.nethealth.com/the-history-of-electronic-health-records-ehrs/#:~:text=The%20history%20of%20EHRs%20began,systems%20to%20adopt%20an%20EHR.&text=In%20the%201960s%2C%20EHRs%20were,in%20partnership%20with%20health%20organizations.> [Accessed 30 January 2022].
[6] Evans, R S. “Electronic Health Records: Then, Now, and in the Future.” *Yearbook of medical informatics* vol. Suppl 1,Suppl 1 S48-61. 20 May. 2016, doi:10.15265/IYS-2016-s006 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5171496/

[7 2022. [online] Available at: <https://www.klearstack.com/information-extraction-in-unstructured-data/> [Accessed 30 January 2022]. https://www.klearstack.com/information-extraction-in-unstructured-data/

[8] Aapc.com. 2022. *CPT - CPT Codes - Current Procedural Terminology - AAPC*. [online] Available at: <https://www.aapc.com/resources/medical-coding/cpt.aspx> [Accessed 30 January 2022].

https://www.aapc.com/resources/medical-coding/cpt.aspx [ht](https://www.cms.gov/Medicare/Fraud-and-Abuse/PhysicianSelfReferral)tps://www.cms.gov/Medicare/Fraud-and-Abuse/PhysicianSelfReferral

[9] Medium. 2022. *Deep learning for specific information extraction from unstructured texts*. [online] Available at: <https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada> [Accessed 30 January 2022]. https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada

[10]https://arxiv.org/pdf/1801.06146.pdf

[11] AI & Machine Learning Blog. 2022. *Named Entity Recognition | NLP with NLTK & spaCy*. [online] Available at: <https://nanonets.com/blog/named-entity-recognition-with-nltk-and-spacy/> [Accessed 30 January 2022].

https://nanonets.com/blog/named-entity-recognition-with-nltk-and-spacy/

[12] AI & Machine Learning Blog. 2022. *Named Entity Recognition | NLP with NLTK & spaCy*. [online] Available at: <https://nanonets.com/blog/named-entity-recognition-with-nltk-and-spacy/#:~:text=Named%20entity%20recognition%20(NER)%20is,person%2C%20location%2C%20organisation%20etc.&text=Building%20a%20highly%20accurate%20NER,%2C%20machine%20learning%20%26%20image%20processing.> [Accessed 30 January 2022].

https://nanonets.com/blog/named-entity-recognition-with-nltk-and-spacy/#:~:text=Named%20entity%20recognition%20(NER)%20is,person%2C%20location%2C%20organisation%20etc.&text=Building%20a%20highly%20accurate%20NER,%2C%20machine%20learning%20%26%20image%20processing.

[13] Analytics Vidhya. 2022. *Named Entity Recognition (NER) in Python with Spacy*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/> [Accessed 30 January 2022]. https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/

[14] Heo, Seok-Jae et al. “Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data.” *International journal of environmental research and public health* vol. 16,2 250. 16 Jan. 2019, doi:10.3390/ijerph16020250

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6352082/

[15] Xu, D., Zhang, M., Zhao, T., Ge, C., Gao, W., Wei, J. and Zhu, K., 2015. Data-Driven Information Extraction from Chinese Electronic Medical Records. *PLOS ONE*, 10(8), p.e0136270. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136270

[16] Yang, J., Liu, Y., Qian, M., Guan, C. and Yuan, X., 2019. Information Extraction from Electronic Medical Records Using Multitask Recurrent Neural Network with Contextual Word Embedding. *Applied Sciences*, 9(18), p.3658. https://www.mdpi.com/2076-3417/9/18/3658

[17]https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0040-1702001

[18] Wu, Yonghui et al. “Clinical Named Entity Recognition Using Deep Learning Models.” *AMIA ... Annual Symposium proceedings. AMIA Symposium* vol. 2017 1812-1819. 16 Apr. 2018 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/

[19] Wu, Yonghui et al. “Clinical Named Entity Recognition Using Deep Learning Models.” *AMIA ... Annual Symposium proceedings. AMIA Symposium* vol. 2017 1812-1819. 16 Apr. 2018 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/

[20] Jouffroy, Jordan et al. “Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study.” *JMIR medical informatics* vol. 9,3 e17934. 16 Mar. 2021, doi:10.2196/17934 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8077811/

[21] Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E., Wu, J., Welt, J., Foote, J., Moseley, E., Grant, D., Tyler, P. and Celi, L., 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLOS ONE*, 13(2), p.e0192360. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192360

[22] Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata, Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, Yasushi Matsumura, “Extracting clinical terms from radiology reports with deep learning”,Journal of Biomedical Informatics,Volume 116,2021,103729,ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2021.103729>. (https://www.sciencedirect.com/science/article/pii/S1532046421000587)https://www.sciencedirect.com/science/article/pii/S1532046421000587

[23] Montagnon, E., Cerny, M., Cadrin-Chênevert, A. *et al.* Deep learning workflow in radiology: a primer. *Insights Imaging* **11,**22 (2020). <https://doi.org/10.1186/s13244-019-0832-5> https://insightsimaging.springeropen.com/articles/10.1186/s13244-019-0832-5

[24] Adam Klie, Brian Y Tsui, Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, Hannah Carter, Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition, Database, Volume 2021, 2021, baab021, <https://doi.org/10.1093/database/baab021> https://academic.oup.com/database/article/doi/10.1093/database/baab021/6259052?login=false

[25] Klie, A., Tsui, B., Mollah, S., Skola, D., Dow, M., Hsu, C. and Carter, H., 2018. Increasing metadata coverage of SRA BioSample entries using deep learning based Named Entity Recognition. https://www.biorxiv.org/content/10.1101/414136v1.full

[26] Textkernel. 2022. *Sequence Labeling via Deep Learning - The magic behind Extract! 4.0 - Textkernel*. [online] Available at: <https://www.textkernel.com/newsroom/sequence-labeling-via-deep-learning-the-magic-behind-extract-4-0/> [Accessed 30 January 2022]. [h](https://www.textkernel.com/newsroom/sequence-labeling-via-deep-learning-the-magic-behind-extract-4-0/)ttps://www.textkernel.com/newsroom/sequence-labeling-via-deep-learning-the-magic-behind-extract-4-0/

[27] Víctor Suárez-Paniagua, Renzo M. Rivera Zavala, Isabel Segura-Bedmar, Paloma Martínez,

A two-stage deep learning approach for extracting entities and relationships from medical texts,Journal of Biomedical Informatics,Volume 99,2019,103285,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2019.103285.

(<https://www.sciencedirect.com/science/article/pii/S1532046419302047>) [https://w](https://www.sciencedirect.com/science/article/pii/S1532046419302047)ww.sciencedirect.com/science/article/pii/S1532046419302047

[28] Nurse documentation and the electronic health record By Janet Pagulayan, MSN, RN-BC; Salim Eltair, MPA, MSN, RN-BC, CCRN, NEA-BC, CPHQ; Kathy Faber RN, MSN,CNL
[https://w](https://www.myamericannurse.com/documentation-electronic-health-record/)ww.myamericannurse.com/documentation-electronic-health-record/

[29] Healthit.gov. 2022. *What is an electronic health record (EHR)? | HealthIT.gov*. [online] Available at: <https://www.healthit.gov/faq/what-electronic-health-record-ehr#:~:text=An%20electronic%20health%20record%20(EHR)%20is%20a%20digital%20version,of%20a%20patient's%20paper%20chart.&text=EHRs%20are%20a%20vital%20part,and%20laboratory%20and%20test%20results> [Accessed 30 January 2022]. [https](https://www.healthit.gov/faq/what-electronic-health-record-ehr#:~:text=An%20electronic%20health%20record%20(EHR)%20is%20a%20digital%20version,of%20a%20patient's%20paper%20chart.&text=EHRs%20are%20a%20vital%20part,and%20laboratory%20and%20test%20results)://www.healthit.gov/faq/what-electronic-health-record-ehr#:~:text=An%20electronic%20health%20record%20(EHR)%20is%20a%20digital%20version,of%20a%20patient's%20paper%20chart.&text=EHRs%20are%20a%20vital%20part,and%20laboratory%20and%20test%20results

[30] Types of EHR :

[https](https://www.wheel.com/companies-blog/types-of-ehr-systems)://www.wheel.com/companies-blog/types-of-ehr-systems

[31] Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition, Addendum 2 [Internet]. Gliklich RE, Leavy MB, Dreyer NA, editors. Rockville (MD): Agency for Healthcare Research and Quality (US); 2019 Oct. [https://ww](https://www.ncbi.nlm.nih.gov/books/NBK551878/)w.ncbi.nlm.nih.gov/books/NBK551878/

[32][https://w](https://www.wheel.com/companies-blog/types-of-ehr-systems)ww.wheel.com/companies-blog/types-of-ehr-systems

[33] Kopp, W., Monti, R., Tamburrini, A. *et al.* Deep learning for genomics using Janggu. *Nat Commun* **11,**3488 (2020). https://doi.org/10.1038/s41467-020-17155-y [https://www](https://www.nature.com/articles/s41467-020-17155-y).nature.com/articles/s41467-020-17155-y

[34] Analytics Vidhya. 2022. *What is ELMo | ELMo For text Classification in Python*. [online] Available at: <https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/> [Accessed 30 January 2022]. [http](https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/)s://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/

[35] T. Tran, T.D. Nguyen, D. Phung, S. Venkatesh

Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)

J. Biomed. Inform., 54 (2015), pp. 96-105, 10.1016/j.jbi.2015.01.012

 [36] 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, Scient. Rep., vol. 6.

[37] Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh Deepr: a convolutional net for medical records IEEE J. Biomed. Health Inform., 21 (1) (2017), pp. 22-30, 10.1109/JBHI.2016.2633963

[38] T. Pham, T. Tran, D. Phung, S. Venkatesh

Predicting healthcare trajectories from medical records: a deep learning approach

J. Biomed. Inform., 69 (2017), pp. 218-229, [10](https://doi.org/10.1016/j.jbi.2017.04.001).1016/j.jbi.2017.04.001

[39] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun

Multi-layer representation learning for medical concepts

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2016), pp. 1495-1504

[40] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor AI: Predicting Clinical Events via Recurrent Neural Networks, in: Proceedings of the 1st Machine Learning for Healthcare Conference, Vol. 56 of Proceedings of Machine Learning Research, PMLR, 2016, pp. 301–318.

[41] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, in: Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.

[42] B.C. Kwon, M.-J. Choi, J.T. Kim, E. Choi, Y.B. Kim, S. Kwon, J. Sun, J. Choo, RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records, IEEE Trans. Visual. Comput. Graph.

[43] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G.E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S.L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N.H. Shah, A.J. Butte, M.D. Howell, C. Cui, G.S. Corrado, J. Dean

Scalable and accurate deep learning with electronic health records

npj Digital Med., 1 (1) (2018), p. 18, [10.103](https://doi.org/10.1038/s41746-018-0029-1)8/s41746-018-0029-1

[44][https:](https://medium.com/mosaix/deep-text-representation-for-sequence-labeling-2f2e605ed9d#:~:text=Sequence%20labeling%20is%20a%20typical,to%20as%20a%20%E2%80%9Ctoken%E2%80%9D.&text=Two%20forms%20of%20sequence%20labeling,of%20Speech%20(POS)%20label%20and)//medium.com/mosaix/deep-text-representation-for-sequence-labeling-2f2e605ed9d#:~:text=Sequence%20labeling%20is%20a%20typical,to%20as%20a%20%E2%80%9Ctoken%E2%80%9D.&text=Two%20forms%20of%20sequence%20labeling,of%20Speech%20(POS)%20label%20and

[45][https](https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/)://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

[46] Saranlita Chotirat, Phayung Meesad, Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning, Heliyon, Volume 7, Issue 10,2021,e08216,ISSN 2405-8440,https://doi.org/10.1016/j.heliyon.2021.e08216.(https://www.sciencedirect.com/science/article/pii/S2405844021023197) [https://ww](https://www.sciencedirect.com/science/article/pii/S2405844021023197)w.sciencedirect.com/science/article/pii/S2405844021023197

[47] Tang, Jie, et al. "Automatic Semantic Annotation Using Machine Learning." The Semantic Web for Knowledge and Data Management, edited by Zongmin Ma and Huaiqing Wang, IGI Global, 2009, pp. 106-150. <https://doi.org/10.4018/978-1-60566-028-8.ch006>

[48] Yousif, M., van Diest, P.J., Laurinavicius, A. et al. Artificial intelligence applied to breast pathology. Virchows Arch 480, 191–209 (2022). https://doi.org/10.1007/s00428-021-03213-3