# Enhancing the Diagnosis of Medical Records to Determine the Clinical Depressions Using ICD-10 Codes

**N. Hema , S. Justus**

***ABSTRACT--- The ICD-10 code provides accurate and updated procedural codes for the improvement of health care diagnosis, cost and ensures an im-partial reimbursement policies. ICD-10-CM is followed and implemented internationally to provide a quality health care for the patients on a global scale. The clinical environment knowledge in a natural language form detects each sentence. In order to maintain positivity, remove all the negative words in the sentence. Dependent clause that provides a sentence element with additional information and which cannot stand alone in a sentence are identified and removed. The resultant sentence is then preprocessed using Text mining techniques. The extracted meaningful words are then processed through the available huge volume of ICD-10 CM codes database. The main aim of this paper is to map the perception of complaints onto an abstract representation and reasoning the system to generate an appropriate ICD-10 CM code. The idea of the work is to provide efficiency on complex vocabulary, vague and imprecise terms, synonymy and polysemy terms. The effectiveness of this proposed work is determined through the process of Perceptron for finding the efficiency between the trained and test dataset.***

***Keywords: ICD-10 CM, ICD-10 PCS, Context Analysis, Text Mining, Stemming, Negation Detection, Business Rule, Global Rule, Perceptron, Knowledge Representation.***

## I. INTRODUCTION

With the introduction of the e-medical record (EMR), the academic medical centres are in need of computer technology to simplify the clinical research [2]. By using the EMR, we are able to maintain large sets of patient record/data beyond the years for study of researchers for identifying patients based on demographic or illness characteristics to address research questions. The data included in the EMR changes based on the hospitals, their common administrative standards and policies, and clinical data fields which may include diagnoses, previous medical histories, current symptoms of the patients, physical examinations, results of tests and procedure undergone, medical diagnoses, referrals to specialists, treatments, medications, and inpatient discharge information [4].

The purpose of this study was to evaluate the Structured or Unstructured EMR data and assign with an equivalent ICD-10 code to identify patients with clinical depression. This standard of diagnosis, using ICD-10 codes may provide a unique code for patient record, which may be helpful for many analyses like “Climatic Disease Identification”, “Geographical Disease Identification” and many. This process of enhancement using ICD-10 code reduces medication error, improves treatment options and disease outcomes and decreases claim submissions. ICD-10 codes are thus shaping the future of clinical practice. The earlier ICD codes are included in practice by the doctors; the better it is for them and their patients.

## II. OVERVIEW OF ICD-10-CM AND ICD-10-PCS CODE

### *2.1. ICD-10-CM*

Worldwide Classification of Disease takes up multi-dimensional reason and utilization, which makes the grouping generally unpredictable. It has turned into the worldwide standard symptomatic characterization for all broad epidemiological & numerous healthcare management purposes. ICD code reduces the need for attachments to explain the patient’s condition and designs a payment system and processing claims for reimbursement. It additionally helps in examination of the general health situation of population gatherings and the checking of the frequency and predominance of sicknesses and other medical issues recorded on numerous health and crucial records [3].

ICD-10 represents International Statistical Classification of Diseases and associated Health Problems. It classifies the data recorded under headings such as "diagnosis", "reason for admission", "conditions treated" and "reason for consultation", which shows up on a wide range of health records through which measurements and other health circumstance data are inferred[3].

In addition, in the present arrangement of ICD code determination, identification of primary term, sub term or increasingly explicit term is exceptionally mind boggling, as it depends in transit the human learning translates. More changes are there to keep running with the risk of missing finding, deferred analysis or wrong determination of the ICD codes.[13]

**Revised Manuscript Received on February 11, 2019.**

**N. Hema,** Assistant Professor, SCSE, VIT University, Chennai. (hema.n@vit.ac.in)

**S. Justus,** Associate Professor, SCSE, VIT University, Chennai. (justus.s@vit.ac.in)

| FORMAT OF ICD-10 DIAGNOSIS CODE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| In this example, S52 is the category. The 4th and 5th characters of "5" and "2" provide additional clinical detail and anatomic site. The 6th character indicates laterality, i.e., right radius. The 7th character, "A", is an extension that provides additional information, which means "initial encounter" | CATEGORY | | | | ETIOLOGY, ANATOMIC SITE, SEVERITY, OR OTHR CLINICAL | | | EXTENSION |
| Characters | 1st | 2nd | 3rd | | 4th | 5th | 6th | 7th |
| Fracture of forearm | S | 5 | 3 | | | | | |
| Fracture of lower end of radius | S | 5 | 3 | . | 5 | | | |
| Torus fracture of lower end of radius | S | 5 | 3 | . | 5 | 2 | | |
| Torus fracture of lower end of right radius | S | 5 | 3 | . | 5 | 2 | 1 | |
| Torus fracture of lower end of right radius, initial encounter for closed fracture | S | 5 | 3 | . | 5 | 2 | 1 | A |

**Fig. 1: ICD-10 CM Format**

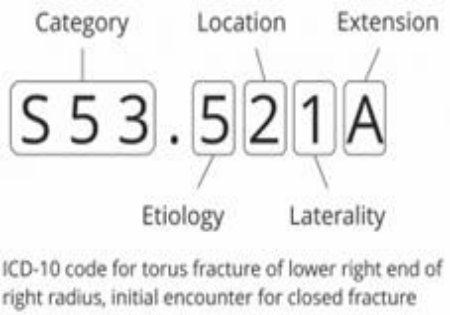


**Fig. 2: Anatomy of an ICD-10 CM code**

*2.2. ICD-10-PCS*

Each ICD-10-PCS code follows a multi-axial code structure which remains descriptive, and expands to represent the details about the procedure used. The procedure coding system (PCS) retrieves the data, gathers the information related to the data, analysis the payment, and maintenance an e-health record for all patient amenities and the respective processes followed. The PCS system of coding provides a seven digit code, where each character of the 7 digit code is given a distinctive attribute that provides a detailed explanation of the medical procedure done to the patient. ICD-10-PCS code has 16 sections ranging from 0 through 9, and alphabets ranging from B through D and F through H. The procedure sections are categorized on a very broad range from surgical procedures to an abuse treatment. These 16 sections are basically classified into three subsections as Medical and Surgical, Medical and Surgical-Related, and Ancillary.

| Character | Character | Character | Character | Character | Character | Character |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Section | Body System | Root Operation | Body Part | Approach | Device | Qualifier |

**Fig. 3: Anatomy of ICD-10-PCS Code**

## III. CODING PROCESS

Coding the patient symptoms to a medical diagnosis code is a complex, difficult and a challenging task, as it involves a lot of procedure and rules to be followed.

*Steps for Correct Coding of ICD-10 CM codes:[14]*

- Analyze the symptoms, signs, diagnosis and conditions based on which to be coded. If symptoms are present but diagnosis remains undetermined, then code the symptoms. The analyzed conditions are not to be coded. They are referred with a keyword "rule out", "suspected" or "questionable".
- Always analyze the code from the Alphabetic Index, before looking into tabular list. To prevent wrong diagnosis of code or error on diagnosis use both the AI and TL when locating and assigning a code.
- Identify the main entry term or the primary diagnosis or main terms, which are written using boldface type. Sub terms and sub term's sub term are also identified.
- Read and interpret any notes listed with the main term like *Excludes1, Excludes 2* and *Includes*. Notes are represented using italicized type.
- Identify and interpret entries for modifiers. Nonessential modifiers are in parenthesis and they do not affect code assignment. Use of parenthesis ( ) means supplementary words are enclosed which follows a diagnostic term without affecting the code number.
- Decipher shortenings, cross reference, images and sections. Cross references utilized are "see", "see additionally", "see class". NEC , NOS, [ ], extra code.
- Recognize a provisional code and discover it in the Tabular List utilizing terms consideration or rejection, notes and different directions, for example, "code first" and " utilize extra code".
- Analyze whether the identified code is at the highest level of specificity. If there is 4th digit applicable, assign a 3 digit code. Assign a 4 digit code if there is no 5th digit code and so on.
- Consult the reimbursement prompts and official guidelines, including the other particulars of the patient.

*3.1. Look up term in Alphabetic Index (AI)*

After receiving the diagnosed statement, identify the main term, sub-term and further classifications of sub-term. The main term is then looked up in the Alphabetic Index and proceeded with the related sub-term symptoms till no more classifications are possible. This leads to exact diagnosis of ICD v 10 Code. If the identified main term is not proper, it may lead to wrong diagnosis.

Example: If the diagnosed statement happens to be as below;

Type I diabetes mellitus with diabetic nephropathy – Here the main term are Diabetes and Type I, and the sub- term is nephropathy.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- type 1 E 10.9
- - with
- - - amyotrophy E10.44
- - - arthropathy NEC E10.618
- - - auytonomic (poly) neuropathy E10.43
- - - cataract E10.36
- - - Charcot's joints E10.6.10
- - - chronic kidney disease E10.22
- - - circulatory complication NEC E10.59
- - - compliucation E10.8
- - - - specified NEC E10.69
- - - dermatitis E10.620
- - - foot ulcer E10.621
- - - gastroparesis E10.43

**Fig. 4(a): Alphabetic Index Representaion for Fever**
Source : http://www.cdc.gov/nchs/icd/icd10.htm

*3.2. Verify code in Tabular Index (TI)*

The diagnosed ICD code in the Alphabetic Index is then verified through a Tabular Index. Taking the ICD code, an equivalent diagnosis statement is retrieved. If this retrieved statement happens to be the same as the statement considered in Alphabetic Index, then the generated ICD code is correct. Else there is a misunderstanding in identifying the main term.

E10 Type 1 diabetes mellitus

E10.1 Type 1 diabetes mellitus with ketoacidosis

E10.10 …… without coma

E10.11 …… with coma

E10.2 Type 1 diabetes mellitus with kidney complications

E10.21 Type 1 diabetes mellitus with diabetic nephropathy

E10.22 Type 1 diabetes mellitus with diabetic chronic kidney disease

E10.29 Type 1 diabetes mellitus with other diabetic kidney complication

E10.3 Type 1 diabetes mellitus with ophthalmic complications

E10.31 Type 1 diabetes mellitus with unspecified diabetic retinopathy

E10.311 …… with macular edema

**Fig. 4(b): Tabular list representation for Fever [19]**

This existing system of generating a solution to the problem is attained through manual interpretation, and specialized knowledge obtained by proper training and guidance. This adds meaning to the diagnosed data. Subsequently there are more chances of execution constraint or quality of the outcome.

## IV. IMPLEMENTATION PLAN

The overview of the proposed system is depicted by the following figure:

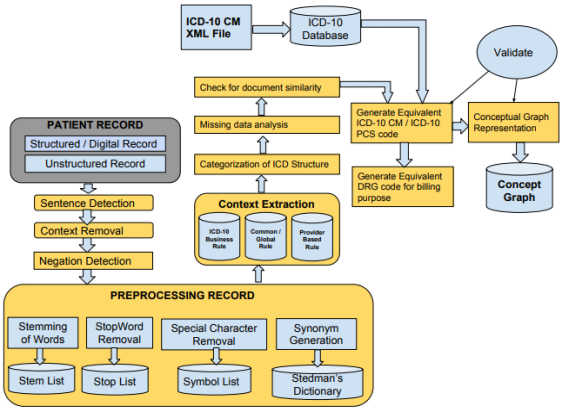


**Fig. 5: Proposed Architecture**

In the above proposed approach, the document can be either in a Structured format such as a defined Patient data Sheet, or an Unstructured format such as an image or a Natural Language file format.

*4.1. Negation Detection*

The purpose of this is to describe a test proposal to capture negative findings in electronic health record systems. A part of the patient observations entered into HER as patient records may have 'negations' or 'negative quantifiers' or 'negative findings' made by the clinicians, for example; statements documented for something *is not the case*. Examples of such negative findings may be like, 'no history of trauma to chest', 'no history of fever and productive cough', or 'no past history of similar symptoms'. These negative findings are as important as positive ones for accurate medical decision-making. They may cause failure to document which may lead to medicolegal results regarding cases of negligence.[9]

A similar query for "not" returns 7272 descriptions, including: "not breathing", "not constipated", "not feeling great", hypertension ruled out", "no significance of hypertension", "no evidence of metastases in the lung", "absence of metastases in the lung", and "abortion was prevented", "kidney not palpable", etc., and for "negative" 1058 descriptions, including: "Joint stress test negative". We do have the not constructors such as, does not , did not , has not , is not , not , was not , not to , had not , do not , could not, should not.

It is also considered as a sentence or phrase that expresses the opposite of a positive statement. The helping verb is usually a form of be, do, or have. We create negative sentences by adding the word 'not' after the auxiliary, or helping, verb. The list of Auxillary words in English are:

| | | | | | | |
|---|---|---|---|---|---|---|
| Do | does | did | has | have | had | is |
| am | are | was | were | be | being | been |
| may | must | might | should | could | would | shall |
| will | can | | | | | |

These Negation Detection are well analyzed through NegEx and applying DEEPEN rules. The DEEPEN rules analyses the presence of;

- Conjunction And (conj_and) Rule : Relation of dependency for a sentence with “conj_and” term.
- Preposition Without (prep_without) Rule : Relation of dependency for a sentence with “prep_without” term.
- Prepositions like prep_in, prep_with, prep_within Rule : Dependency connection for a sentence with “prep_with” relation.
- Nominal Subject (nsubj) Rule : Relation of dependency for a sentence with “nsubj” relation.
- Suggest Rule : Relation of dependency for a given sentence with “suggest” as the term of dependency.

### *4.2. Pre-Processing*

Pre-processing is the most important part of mining a text application. The given structured format is preprocessed in 4 stages as; Stemming, Stop Word removal, Synonym Generation and Conditions, Signs and Symptom identification.

#### *4.2.1. Stop Word Removal*

Stop words are natural language words. They are the words that carry no importance in analysis, and are removed to make the document look much simple. The articles, pronouns, prepositions etc., adds no meaning for the document, and hence are added into the Stopword list[2].

Stopword removal is done by :

- Creating a file with a list of stopwords as[18];

| |
|---|
| an about above after again against all am an and any are aren't as at be on the grounds that been before being underneath between both yet by can't could couldn't did didn't do does doesn't doing don't down amid every few for from further had hadn't has hasn't have haven't having he he'd he'll he's her here's hers herself him himself his how's I i'd i'll i'm i've if in into is isn't it's its itself allows me all the more most mustn't my myself no nor not of off on once just or …….. |

**TABLE I - SAMPLE STOPWORDS**

- Tokenizing the individuals in the above input document.
- Read a single stop word from the stopword list file and compare with the tokenized word using sequential search technique and are removed from the tokens.
- If it matches, the word in the array is removed. This is done for the entire array list.

#### *4.2.2. Stemming*

The Stemming process provides ways for finding morphological variants of search terms. This process improves the retrieval effectiveness and reduces the size of indexing files[15].

There are many types of stemming algorithms like;

- Porter’s algorithm
- Suffix stripping algorithm
- Lemmatisation algorithm
- Hybrid approach

The performance of the stemmer is assessed by examining its behaviour on a samples of words that are already arranged into a 'conflation groups'. This allows, specific errors that fail to merge or wrongly merge (e.g., "maintained" with "maintenance", and "experiment" with "experience") to be identified and adjusted.

The word paining, painful, pains, pained can be stemmed to a word ‘pain’. This is done to reduce the number of words in the document and to have accurate meaning, thereby reducing time and saving memory.

For example: The clinical stemming words are like;

| | |
|---|---|
| Diabetes | Diabetic |
| Low Carbohydrate diet | Low carb diet |
| Blood glucose | Blood sugar |
| Type 2 diabetes | Diabetes mellitus |
| Blood sugar | Haemoglobin A1c |
| Appetite | Appetit |
| Syndrome | Syndrom |
| Medication | Medic |

**TABLE II - STEMMING WORDS**

#### *4.2.3. Synonym Generation*

Synonym generation as far as medical terminologies are concerned, plays a very important role. A word may be referred using different terms with same synonym by various people. We need a technique to identify all such similar words, and to derive at a common word, than many words holding the same meaning. This is done to get better analysis result. It also reduces the number of words in the document, thereby reducing memory and time.

Example: The words fever, temperature, high temperature, running high temperature etc., all refers to the same meaning as ‘fever’.

This synonym generation is done using Lexicon Synonyms.

*Lexical synonyms:*

These are grammatical synonyms as defined by the rules of the language. WordNet is a popular source for these among others. Lexical simplification is a form of text simplification that focuses from the word level. It is performed by the process of words substitution using simpler synonyms, adding definition, or by showing simpler synonyms.For Example:

| **Word** | **Embedding words** |
|---|---|
| Nausea | vomit, vomiting, sickness, queasiness |
| Swell | puffiness, lump, swell up |

Also, using Stedman's database lexical synonym can be substituted with relevant words as;

| **Word** | **Embedding words** |
|---|---|
| Biovar cholera | Cholera Classical, Cholera due to vibrio c cholerae o1 |
| Biovar eltor | Cholera eltor |
| Typhoid Pneumonia | Pneumonia in typhoid fever |
| Typhoid Arthritis | Arthritis caused by salmonella typhosa |
| Hordeolum internum | left upper eyelid, L upper eyelid meibomitis |

Even at this stage, it is noted that there are many words which add no meaning for the process. These irrelevant words can be removed by intersecting them with the Tailored words of ICD-10-CM codes[2].

The tailored words are obtained from the Alphabetic Index table of ICD-10 (*https://www.cihi.ca/en/icd_volume_two_2012_ en.pdf).* The words in figure are now intersected with the Alphabetic Index table of ICD-10.

*4.2.4. Conditions, Signs and Symptoms Identification*

Codes that show the side effects and signs, instead of findings, are satisfactory for reporting purposed when a related absolute analysis has not been set up by the provider. Some of them will be transient and never need investigation. While other need investigation but reveal nothing. Still some of them may be signs and symptoms of underlying diseases which may be major or minor. ICD-10 CM codes from R00.0 to R99 contain many codes for symptoms, which are not coded when they are related to the diagnosis. Sometimes it is possible to code these symptoms, because uncertain diseases cannot be coded.

*4.3. Context based information extraction*

A relationship extraction is the process of detection and classification of semantic relationship within a set of artifacts, typically from text or XML documents. Extraction of information based on the context may help the clinical information more accessible. This task is like the data extraction (IE), yet requires the deletion of repeated relations without confusion and generally refers to the extraction of a wide range of different relationships. Extraction of relationship of the preprocessed words is done to identify the dependency, inclusions and exclusions of the words. This phase helps to analyze the root nodes and the sub-nodes, thereby grouping the diagnosed diseases together.[7]

*4.3.1. Business Rule*

There are certain conditions existing among ICD-10 codes on combination of "use additional code" and a "code first" note in the medical record.

- The "Code First" node is used as the primary code in the combination, which requires a manifestation code in bracket to be followed.
- Code Also - Guidelines that tells the coder that beyond what one code could be appointed, but it doesn't infer any sequencing direction. For the most serious conditions have to be recorded first.
- The "Use additional code" is used as an secondary code, that is a part of manifestation combination.
- No coding for document like "probable", "suspected", "rule out", "working diagnosis" and "questionable".
- The keyword "Includes" specifies that there are further definitions or examples on the content of the category.
- "See" or "See Also" term used to instruct the coder to refer to another term if desired.
- "Excludes 1" indicates that the code should never be used along with the code above the Excludes 1 note. It may be helpful to think of the "Excludes1" list to be codes that might be suggested instead.
- "Excludes 2" indicates that the excluded condition is not part of the code required condition, but a patient can have both. It may be helpful to think of "Excludes 2" codes to those that might need to be added to provide full detail.
- The "With" of "Associated with" or "due to" term is used immediately following the main term. If not specified in the documentation, the default is "without".

*4.3.2. Global Rule*

Association rule is a rule-based machine learning method for discovering clinical findings and interesting relations between variables or hidden patterns in the clinical databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. An association rule is expressed in the form X→ Y, with X referring the antecedent of the rule and Y referring the consequent. These rules may also be nested, if the RHS of the both the rules are same and the LHS of one rule is a subset of the LHS of the other rule[1]. Association rules are widely used for finding correlations within terms in textual data. The strength of the formed association rules are measured using Apriori algorithm, FP Growth, Eclat, Apriori TID, RElim and Close algorithms[12].


*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

| |
|---|
| Accupril → Hypertension |
| Insulin → Diabetes |
| Lantus → Diabetes |
| Glucophage → Diabetes |
| Lipitor → Hyperlipidemia |

**TABLE III - RULES OF THE FORM MEDICINE → DISEASE**

| |
|---|
| Diabetes → Hypertension |
| Congestive heart failure → Diabetes |
| Hypertension → Diabetes |
| Hyperlipidemia → Diabetes |
| Hypertension → Hyperlipidemia |
| Acute renal failure → Diabetes |

**TABLE IV - RULES OF THE FORM DISEASE→ DISEASE**

| |
|---|
| Anemia → Diabetes |
| Chest pain → Congestive heart failure |
| Nausea → Hypertension |
| Nausea → Diabetes |
| Obesity → Diabetes |

**TABLE V - RULES OF THE FORM SYMPTOM → DISEASE**

*4.4. Categorization of ICD-10 Structure*

The pre-processed data are organized to a specific ICD-10 component such as Category, Ethiology, Severity, Laterality and Extension. This helps in removal of irrelevant data. The relevant data are then organized based on their dependencies. The organized information's are later analyzed for missing data in any one or two of the ICD-10 structured component. These missing values are substituted through previous information available or through a Provider Based Rule[16].

The final structured information's available on patient record are compared for similarity with already existing record for easy ICD-10 code generation to save time and effort. This comparison and replacement of equivalent ICD-10 code is done using K-means algorithm. There are possibilities for one patient record to have an equivalent replacement from two or three previous patient records. This is done using Fuzzy K-means algorithm.

*4.5. Conceptual Graph Representation*

Knowledge begins with limited data. Organizing and analyzing the data gives the required information about the data. This obtained information is termed as 'knowledge', and can be used represented information about the patient in many ways. This information represented is used to solve complex problems, to make inferences, give reason about the knowledge and obtain new knowledge [11]. Each patient is denoted by a node and all their medical treatments are represented as sub nodes with the date of visit for future querying.[11]

## V. PERFORMANCE METRICS USING PERCEPTRON

The extracted knowledge as described in the proposed system can be compared using Simple Perceptron a learning algorithm a deep learning process. Perceptron provides a better understanding of Neural Nets, which is a binary classification algorithm. It takes in numerical inputs and produces a single binary output.[6]

The 2 inputs to treat a perceptron are;

- Input data (numerical)
- Weight for each input data (numerical) which describes the im portance of that input.

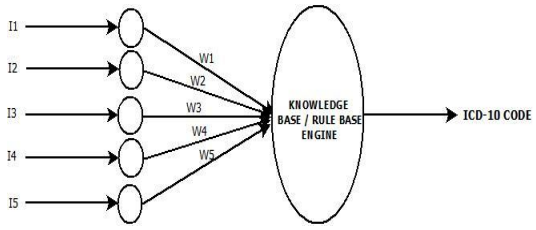


**Fig. 6: Simple perceptron**

This Neural Network algorithm helps to learn the weights from training data, which can be used to predict future unseen data, and performs a Weighted Sum. In the process of Weighted Sum or Summation phase, the input data along with their weights get summed up. If the input data holds a vector [I1, I2, I3, I4, I5], and its corresponding weights are [w1, w2, w3, w4, w5], then the weighted sum is the dot-product of the inputs along with its weights.

$$\text{Weightedsum} = (\sum \text{iwi.Xi}) = \text{wT.X}$$

The five input data are assumed to be the Gender, Category or the Core code, Etiology, Anatomic Site, Severity and the Extension. The obtained weighted sum is activated or fired based on the

Fuzzy rules existing in the knowledge base. The received output if then compared with our actual output, and the error is calculated using a loss function and weighted sum.

The Fuzzy Rules may be of the type:

*If (Gender is Male) and (Category is Tuberculosis) and (Etiology is Genitourinary System or Etiology is Urogenital System) and (Anatomic Site is Cervix) then (ICD-10 Code is Error)*

*If (Gender is FeMale) and (Category is Tuberculosis) and (Etiology is Genitourinary System or Etiology is Urogenital System) and (Anatomic Site is Cervix) then (ICD-10 Code is A18.16)*

*If (Gender is Male or Gender is Female) and (Category is Tuberculosis) and (Etiology is Bones and Joints) and (Anatomic Site is Spine) then (ICD-10 Code is A18.01)*

*If (Gender is Male or Gender is FeMale) and (Category is Tuberculosis) and (Etiology is Genitourinary System or Etiology is Urogenital System) and (Anatomic Site is Bladder) then (ICD-10 Code is A18.12)*
*If (Gender is Male or Gender is FeMale) and (Category is Tuberculosis) and (Etiology is Genitourinary System or Etiology is Urogenital System) and (Anatomic Site is Kidney and Ureter) then (ICD-10 Code is A18.11)*

The performance of the classifier [1] is measured using :

*5.1. Classification Accuracy*

The classification accuracy Ai of an individual program i depends on the number of samples correctly classified (true positives plus true negatives) and is evaluated by the formula:

$$Ai = ( t / n ) * 100 \%$$

where "t" is the number of sample cases correctly classified, and "n" is the total number of sample cases.
For example, considering n=150 number of sample patient cases, we have obtained t=130 number of correctly classified cases. The accuracy Ai is said to be 86.67%

5.1.1. Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a given set of test cases. It is the average over the test sample of the absolute differences among expectation and actual perception where every single individual distinction have equal weight.

$$MAE = (|a1 – c1| + |a2 – c2| + … +|an – cn|) / n$$

Here 'a' is the actual output and "c" is the expected output. The prediction "yes" is assigned a weightage of 1 and "no" is assigned a weightage of 0 respectively.

*5.1.2. Root Mean Square Error (RMSE)*

RMSE is a quadratic scoring rule that estimates the normal mag-nitude of the mistake. It is the square base of the normal of squared contrasts among predicted and real perception. It is calculated as:

$$RMSE = [\sqrt{((a1 – c1)2 + (a2 – c2)2+ … + (an – cn)2 )}] / n$$

Here 'a' is the actual output and 'c' is the expected output. The prediction "yes" is assigned a weightage of 1 and "no" is assigned a weightage of 0 respectively. The mean-squared error is the commonly used measure for numeric prediction.

*5.1.3. Confusion Matrix & Results*

A Confusion Matrix is regularly used to portray the execution of a classification model (or "classifier") on identified test data information for which the true or positive qualities are known.

| | Predicted 1 | Predicted 0 |
|---|---|---|
| True 1 | True Positive | False Negative |
| True 0 | False Positive | True Negative |

**Fig. 7: Confusion Matrix**

| N=150 | Predicted : No | Predicted : Yes |
|---|---|---|
| Actual : No | 25 | 13 |
| Actual : Yes | 7 | 105 |

**Fig.8: Confusion Matrix for 150 Patient records**

Accuracy = correct / (correct + incorrect)
= (25 + 105) / (25 + 7 + 13 + 105)

In the above Confusion Matrix, there are two classes of prediction: "yes" and "no". Here, "yes" means the positive prediction of the presence of disease, and "no" mean absence of the disease with the patient.

On a total, the classifier has made predictions on 150 patient for the positive presence of the disease.

Of the 150 cases, there are 118 positive predictions (yes) and 32 negative (no) predictions identified.

But in reality, 112 patients from the sample have the disease, and 38 patients do not have the disease.

The negative (no) predictions of 7 records by the proposed system may be due to some Local Rule adopted by the physician or by the respective hospital.

*5.1.4. Precision and Recall*

Precision is the fraction of retrieved documents that are relevant to the query. It is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate.

$$\text{Precision} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Positive})}$$

For the above Confusion Matrix, precision is calculated as,

Precision = 25 /(25 + 7)
= 0.78

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved. It is defined as the ratio of correctly predicted positive observations to all observations in actual class - yes.

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$$

For the above Confusion Matrix, recall is calculated as,

Recall = 25 /(25 + 13)
= 0.66


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

## VI. CONCLUSION

This paper has proposed a new form of representing the digital medical records to represent a common ICD -10 code which may be helpful for analysis of many things. This work was helpful in carrying out a informative study about digital health record, its necessity for a code conversion and the benefits of cloud storage. By deploying this paper, it can be made possible that the proposed system would be functioning as expected and the results to be proven through Deep Learning process. It was helpful in predicting the type of disease with few keywords unaware of the underlying processes carried out in efficacious predictions. The new system can be a vantage to the field of medical studies.

## REFERENCES


1. MingkaiPeng, VijayaSundararajan, TylerWilliamson, Evan P.Minty, Tony C.Smith, ChelseaT.A.Doktorchik, HudeQuan, DATA ON CODING ASSOCIATION RULES FROM AN INPATIENT ADMINISTRATIVE HEALTH DATA CODED BY INTERNATIONAL CLASSIFICATION OF DISEASE - 10TH REVISION (ICD-10) CODES. *Elsevier Inc., February 2018; https://doi.org/10.1016/j.jbi.2018.02.001, Data in Brief 18 (2018) 710–712*
2. Ricardo Baeza-Yates, Luz Rello, Julia Dembowski. CASSA: A Context-Aware Synonym Simplification Algorithm. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1380–1385, Denver, Colorado. 2015 May – 2015 June, Association for Computational Linguistics
3. Jaideepsinh K. Raulji, Jatinderkumar R. Saini. Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. *International Journal of Computer Applications.* 2016 September; 150(2);0975 – 8887.
4. N. Hema, S. Justus. Conceptual Graph Representation Framework for ICD-10. *2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Elsevier Procedia Computer Science* 50 ( 2015 ) 635 – 642.
5. N. Hema, S. Justus. Knowledge Base Representation Model for a Structured Repository. *International Journal of Engineering Science and Computing*. 2016 March, DOI 10.4010/2016.515, ISSN 2321 3361 © 2016 IJESC
6. Lakshmi Devasena C. Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction. International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 4, April 2014
7. Lakshmi K.S, G.Vadivub. Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.
8. Nhi-Ha T. Trinh, Soo Jeong Youn, Jessica Sousa, Susan Regan, C. Andres Bedoya, Trina E. Chang, Maurizio Fava, and Albert Yeung. Using Electronic Medical Records to Determine the Diagnosis of Clinical Depression. Int J Med Inform. 2011 Jul; 80(7): 533–540.
9. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, Beesley C, Dexter P, Max Schmidt C, Liu H, Palakal M. DEEPEN:
10. A negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform. 2015 Apr;54:213-9.
11. doi: 10.1016/j.jbi.2015.02.010. Epub 2015 Mar 16.
12. J. F. Sowa. Conceptual Structures-- Information processing in mind and machine. Addison-Wesley Systems Programming Series Reading, MA, 1984
13. Ronald J. Brachman, Hector Levesque. Knowledge Representation and Reasoning. Edition 2010.
14. Lakshmi Devasena C, ICD-10 Volume 2 Instruction Manual 2010 Edition.
15. Erin P. Balogh, Bryan T. Miller, and John R. Ball, Editors; Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine; The National Academies of Sciences, Engineering, and Medicine
16. Article title. https://www.cms.gov/Medicare/Coding /ICD10 /.../ ICD10Introduction.pdf. Date accessed: 05/01/2015
17. Article title. https://en.wikipedia.org/wiki/ Stemming. Date accessed: 07/03/2015
18. Article title. www.textanalysis.com. Date accessed: 15/10/2015
19. Article title. http://www.icd10data.com/ ICD10CM/Codes. Date accessed:22/03/2016
20. Article title. http://www.ranks.nl/stopwords. Date accessed: 10/06/2016
21. Article title. http://www.icd10data.com: 10/06/2016


## AUTHOR PROFILE

**N. Hema** is an Assistant Professor (Senior) from the School of Computing Science and Engineering at Vellore Institute of Technology, Chennai. Her PhD is on Medical Diagnosis using ICD-10 Codes and representation of the diagnosed knowledge using Conceptual Graph.

**Dr. S. Justus** received his doctorate degree from Madurai Kamaraj University, Madurai, India. His research specializations include Object-relational data modeling, knowledge engineering and Big Data.

He has been into academic research and has published several of his research work results in International Journals and Conferences – including SwSTE in Israel and DASMA in Germany. He has also practitioner's experience while working with Software development companies in India.

He is a member of IEEE, ISTE, IAENG professional associations. He has served as research & project coordinator for PG studies at Engineering Institutes. Presently he is working as Associate Professor and the former chair of Software Engineering Research Group at VIT University, Chennai. He is the co-founder of a software consulting company InClefs through which the team designs and develops software solutions for various clients across the state.