Processing of Clinical Notes for Efficient Diagnosis with Dual LSTM

Chandru A. S[1]
Research Scholar, Department of CSE, Visvesvaraya Technological University (VTURRC), Karnataka, India

Seetharam K[2]
Professor, Department of CSE
Jnanavikas Institute of Technology, Bangalore, India

***Abstract*—Clinical records contain patient information such as laboratory values, doctor notes, or medications. However, clinical notes are underutilized because notes are complex, high-dimensional, and sparse. However, these clinical records may play an essential role in modeling clinical decision support systems. The study aimed to develop an effective predictive learning model that can process these sparse data and extract useful information to benefit the clinical decision support system for the effective diagnosis. The proposed system conducts phase-wise data modeling, and suitable text data treatment is carried out for data preparation. The study further utilized the Natural Language Processing (NLP) mechanism where word2vec with Autoencoder is used as a clustering scheme for the topic modeling. Another significant contribution of the proposed work is that a novel approach of learning mechanism is devised by integrating Long Short Term Memory (LSTM) and Convolution Neural Network (CNN) to learn the inter-dependencies of the data sequences to predict diagnosis and patient testimony as output for the clinical decision. The development of the proposed system is carried out using the Python programming language. The study outcome based on the comparative analysis exhibits the effectiveness of the proposed method.**

Keywords*—*Clinical notes; natural language processing; diagnosis; long short term memory; convolution neural network; autoencoder

I. Introduction

The diagnosis is a critical part of the healthcare system that decides the kind of treatment that needs to be given to the patient and builds the entire treatment strategy. Initial assessment in the diagnosis is a crucial step, and if it goes wrong, it will lead to lots of consequences. One interesting research study has reported that 65% of the medical mishaps are due to wrong diagnosis only and 11% of these cases result in death [1-2] Therefore, developing intelligent models is essential to reduce false diagnoses and to help experts make the right decisions for patient treatment and well-being. However, building efficient diagnostic systems requires the availability of relevant data and predictive models for real-time deployment. Recent advances in Machine Learning (ML) technologies have brought opportunities to improve healthcare and enhance patient outcomes. Furthermore, clinical records are a decent resource that provides a crucial scheme and scope of optimizing the diagnostic process. The clinical notes are nothing but the text written by the physician and the medical experts who have admitted and treated the patient. These clinical notes usually have two parts to it, namely, i) Patient testimony ii) Doctor's notes. These two contain different types of information, which acts as a powerful resource providing detailed patient conditions and clinical inference, which usually cannot be obtained from the other components of the electronic health record [3-4]. These two parts of the clinical notes have distinct importance in order to make diagnosis better. Based on this clinical information, few studies have shown that among the patients who are entering the hospital from the Emergency Room (ER), 29% of them are unconscious and 40% of them have some type of psychological episodes [5-6]. Among them, the patients who needed Cardiopulmonary Resuscitation (CPR) during the admission to hospital via ER, only 11% survived till the discharge [4]. Among the patients who enter the hospital from Out Patient Department (OPD), it is shown that 80% of the patient (Non-psychological conditions) testimony is fully reliable [7-8]. This proves the point that initial diagnosis is the critical step in the healthcare cycle. Hence depending on the entry point of the patient, the system needs to be designed in such a way that it gives higher importance to patient testimony if the patient is entering from the OPD, and gives lower importance to the same while the patient is entering from ER [9-10]. Recent advances in technologies have shown that natural language processing (NLP) and ML algorithms can be used to build an effective Clinical Decision Support (CDS) system to benefit a successful diagnosis with high scores. However, most of the existing works on CDS have employed standard ML algorithms. On the other hand, the clinical notes contain sparse information, abbreviations, unusual grammatical structures and are high-dimensional. Creating models that learn useful representations of clinical texts is a challenge. While ML algorithms learn without human intervention, preparing the suitable data for ML algorithms needs the right algorithm and tuning it for optimal results. In general, the linguistics of these clinical data requires distinct modeling because they contain a degree of ambiguity that requires the use of different approaches and multiple efforts to come up with the most efficient solutions. Therefore, this paper intends to suggest an effective diagnostic system based on the joint approach of NLP and deep learning techniques. The main contribution of the proposed study is highlighted as follows:

- To emphasize the usage of discharge summary of a patient in order to extract more information data associated with admission of patient for leveraging diagnosis.

- To develop an NLP model for facilitating a unique diagnostic process on the basis of clinical notes of patient which could improve accuracy to next level.

- To deploy Autoencoder for carrying out clustering of text in medical dataset that can classify between doctors notes and clinical testimony of patient.
- Further, novel and efficient Dual-LSTM is built that reveals a different importance to the clinical text depending on the result of the text clustering. Basically, it has two levels of importance to the patient's testimony and Doctor's notes and importance may vary depending on where the patient has entered the hospital from.

The prime motivation of the proposed study is to harness the strength of machine learning approach in a unique fashion in order to carry out reliable diagnostic of the disease. At present, the diagnosis of the disease is mainly carried out considering the medical report of a doctor, whereas various information contextual information could be present. Therefore, this could be further enriched if the information is further provided in the form of patient testimony. Hence, a system is developed in such a way that given the data, of various tests and testimony of both patient and doctors, the system can diagnose the patient.

The remaining part of this paper is structured as follows: Section-II presents a brief review of previous research works; Section III discusses the system design and dataset; Section IV elaborates on the implementation procedure adopted in the proposed system; further result and performance analysis is presented in Section V and finally, Section VI concludes the entire effort and findings.

II. Related Work

This section presents a brief review of literature in context of modeling clinical decision support system based on the machine learning technique and clinical data.

A recent work done by Mustafa and Azghadi [11] provided a detailed review study on applying ML techniques for clinical notes in the healthcare industry. The authors have also discussed potential challenges in working with clinical data and highlight open research issue. The authors have also discussed the concept of AutoML and highlights its benefits for processing clinical notes. However, a data treatment operation is effective in predictive modeling performance, especially when dealing with complicated data like clinical data. The research work in the direction of treating clinical notes is carried out by the Kaur et al. [12] suggested a rule-oriented technique for correcting clinical text data. The authors have applied word correction rule to recognize the term and its definitions. Further supervised ML classifier support vector machine is applied to give suitable treatment for cleaning clinical data. This study has demonstrated effectiveness of combination of the rule-based and ML technique in working with the clinical data. The work or Hassler et al. [13] has shown the importance of the clinical data treatment in the predictive modelling. The first step is data preparation based on the statistical and semantic analysis and new features are then extracted. Further, imputation is done to handle the missing data using ML technique. Another work in the similar research line is done by Kashima et al. [14], where a comprehensive analysis is made regarding the impact of preprocessing at every stage of classification process. The authors have done removal of stop words, lemmatization, normalization and stemming for modelling text data which is then vectorized using Bag of words. Further, a logistic regression approach is used in the classification phase. The study outcome claims that normalization and error correction have a highly positive impact on the classifier's performance. Ferrao et al. [15] provided a roadmap to handle complex medical data using preprocessing techniques. In this study, a phase wise data handling operation is shown that includes error identification, treating missing and redundant data, feature analysis, and information retrieval process. In the work of Mishra and Yadav [16] the preprocessing operation over medical data includes k-means imputation, transformation of discreate value, normalization, random forest-based feature selection. In this study the authors have implemented various ML techniques which achieved higher accuracy with preprocessed data. Once the data are treated and cleaned, their features need to be analyzed and selected. However, the above discussed approaches are specifically focused on the preprocessing operations, and adopted standard approach of feature selection. They have not emphasized effective feature modelling from the perspective of the feature engineering problem, which is an important concern in the predictive modelling, its validation and acceptance in the clinical industry. Although there are few research works in this direction with extensive feature engineering, they need optimization or customization in the design of learning models. Teo et al. [17] attempt to predict the chances of patient readmission in hospital using various ML technique. The outcome illustrates complex deep learning technique outperforms shallow ML models. The work of Spasic and Nenadic [18] gave systematic evidence on the performance of ML model trained on the clinical text. The authors have examined the variety of NLP operations supported by ML techniques. The work of Ye et al. [19] employed unified medical language system and Convolutional Neural Network (CNN) to predict the mortality rate. In this work, the authors have used mimic-III dataset which is applied with concept unique identifiers and entity embedding for the textual feature representation. The usage of unified medical language system with ML-NLP is also seen in the study of Weng et al. [20] for the classification of medical sub-domain. Apart from this, Metathesaurus, are Semantic Network used to extract features which is then combined for the classification process using supervised learning. Kumar et al. [21] developed a classification model to predict whether a morbidity occurs for an individual by analyzing his/her medical records. This study utilized pre-trained word2vec, GloVe, fastText, and sentence encoder embeddings for the classification. Topaz et al. [22] performed comparative analysis between the rule-oriented approaches and NLP-based ML techniques for fall detection from the clinical notes. The work done by Poul et al. [23] has developed a linguistics-driven framework using genetic programming to predict the risk of suicide from the analysis of the clinical data. Using clinical notes, Huang et al. [24] applied bidirectional encoder representations from transformers for forecasting hospital readmission. Prabhakar and Won [25] presented hybrid learning model for medical text data classification. The hybrid model is presented based on hybrid Long Short-Term Memory (LSTM) A bidirectional gated recurrent unit is implemented to reduce the human effort in modelling data and feature selection. Moqurrab et al. [26] combined CNN, BI-LSTM and discriminant model for the extraction of the clinical entities from the medical notes. Detecting clinical entities accurately can be helpful in maintaining the confidentiality of medical data, which increases trust between users and medical organizations. Table I other significant learning approaches in the context of processing clinical notes for efficient diagnosis.

Hence, there is much research work being carried out in the literature for modelling clinical notes and predictive system for clinical decision support system. Despite numerous works, there is significant issues associated with effective data processing, and predictive modelling. These significant issues are briefly highlighted as follows:

- It has been analyzed that most of the existing approaches have adopted common mechanism for data preprocessing such as normalization and stemming.
- The previous studies have not carried exploratory analysis to understand the nature of dataset and requirement of preprocessing operations.
- Lack of novelty in the design of machine learning model. The existing approaches do not emphasize the modelling of effective learning systems specific to the clinical data.

Therefore, the problem statement for the proposed system can be expressed as "it is quite challenging to design an optimal predictive model along with better treatment operation on the unstructured, and sparse clinical notes for an effective diagnosis support system"

TABLE I. SUMMARY OF THE EXISTING LITERATURE

Citation	Context	Clinical notes processing	Predictive Modelling
[27]	ICD coding	TF-IDF	CNN and Decision Tree
[28]	Detection of anastomosis leakage	BoW	SVM
[29]	Performance of Predictive model	BoW, Word2Vec	Linear regression (LR) and K-nearest neighbor
[30]	Classification of Medical codes	Glove	LR, CNN, LSTM
[31]	Extraction of medication and adverse drug event	Word2Vec	SVM, LSTM, CNN
[32]	Clinical coding analysis	Word2Vec, Glove	SVM, LR, RF, CNN, LSTM
[33]	Classification of diagnosis codes in discharge notes.	Glove	RF, SVM, CNN
[34]	Feature engineering	cTakes	SVM
[35]	Knowledge extraction	cTakes	SVM, KNN
[36]	automated ICD coding	Word2Vec	Deep neural network

III. SYSTEM DESIGN

The development of the system is done using deep learning techniques, which combines mechanism of both LSTM and CNN architecture. The proposed system also makes use of both NLP (Word2Vec) with autoencoder algorithms to make the final diagnosis. The system is set up as a classification learning model where ICD-9 Codes of the diseases are considered as classes. For this purpose, MIMIC-III [9] dataset has been used, which is collected by Beth Israel Deaconess Medical Center. The schematic architecture of the proposed system is shown in figure 1.

The bock diagram of the entire system is as shown in the figure 1. There are two types of data as it can be observed discharge summary and admission type. The admission type is given to train the dual LSTM (LSTM+CNN). Therefore, proposed dual LSTM knows which data it should give importance to depending on ADMISSION_TYPE. The response of this system is diagnosis i.e., the model outputs the diagnosis based on given data. The diagnosis is encoded in the output with one hot encoding technique. Also, a mechanism of NLP i.e., Word2vec is used to identify whether the data is clinical notes or patient testimony, which is achieved by text clustering operation by the proposed dual LSTM. This is discussed in detail in section 3.5. Hence ultimately, the dual-LSTM mentioned over here is most suited for processing medical data which is the novel contribution of this study.

A. Dataset

The MIMIC-III dataset contains total of 26 tables which contains the data of over 40,000 Patients with their personally identifiable information (PII) is deidentified. Hence even though dates provided in the data are wrong, it is made sure that vital data like age of the patient during admission and the number of days stayed in the hospital are not being changed. The deidentification of PII is done in order to protect the patient privacy.

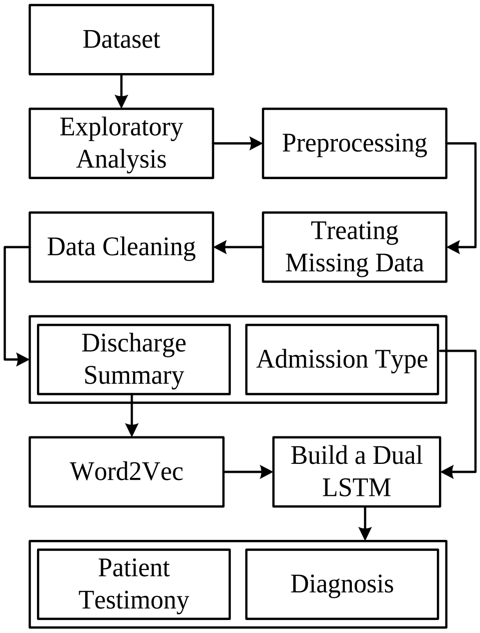

Fig. 1. Block Diagram of the Proposed System.

B. *Scope of the Study*

This study is limited to non-psychological patients only hence the ICD-9 codes which are between 290-319 are rejected since they represent mental disorder. Even though the dataset contains several other medical records, only discharge summary is being considered.

IV. Methodology Implementation

This section discusses the methodology adopted in the proposed system for the prediction of diagnosis and patient testimony.

A. *Exploratory Analysis*

In this phase of study, the dataset adopted is analyzed and it is found that it consists of 26 tables. Among the 26 tables, only 4 are important to the present study and they are, i) NOTEEVENTS.csv ii) DIAGNOSES_ICD.csv iii) D_ICD_DIAGNOSIS.csv and iv) ADMISSIONS.csv. However, the in the study not all the columns in these tables are being considered. Only a select few columns which are necessary for this study are being considered here.

NOTEEVENTS (α):

This table contains important notes written by various therapists and nurses during the patients' hospital stay. Following columns are considered in this table.

- SUBJECT_ID: A unique identifier for the patient.
- HADM_ID: A unique ID given to hospital stay. A single SUBJECT_ID has many HADM_IDs. Used as foreign key.
- CHARTTIME: Date and time at which the note was charted.
- CATEGORY: Type of note.
- ISREEOR: If there is an error with the note and needs repetition.
- TEXT: content of the note.

DIAGNOSIS_ICD (β):

This table contains the patient's final diagnosis in form of ICD-9 codes. Following columns are being considered.

- SUBJECT_ID: A unique identifier to each patient.
- HADM_ID: A unique ID given to hospital stay. A single SUBJECT_ID has many HADM_IDs. Used as foreign key.
- ICD9_CODE: Diagnosis made for the admission.

D_DIAGNOSIS_ICD (γ):

This table contains mapping of the ICD 9 code to name of the disease.

- ICD9_CODE: ICD9 code.
- SHORT_TITLE: name of the disease.

ADMISSIONS (θ):

This table consists of the information about the admission of the patient. Following columns are being considered.

- HADM_ID: used as primary key in this case.
- ADMITTIME: time of admission.
- DISCHTIME: time of discharge.
- ADMISSION_TYPE: ER or OPD.
- DEATHTIME: time of death of patient (If died during hospital stay else NaN).

B. *Preprocessing*

In the preprocessing step all the four tables are initially joined using inner join function. The new table is called as master_data_table (**Ω**) fiven as follows:

$$\mathbf{\Omega} = \theta[\text{HADM\_ID}] \bowtie \alpha[\text{HADM\_ID}] \bowtie \beta[\text{ICD9\_CODE}] \bowtie \gamma[\text{ICD9\_CODE}].$$

In **Ω**, two columns are then removed since they are primary keys for joining the tables. And they no longer hold any significance. They are, i) SUBJECT_ID and ii) HADM_ID. The ICD9_CODE is truncated to first 3 characters. First 3 characters of the ICD9 code always represents a class of disease rather than the full condition. The algorithm is optimized to recognize the class of disease rather than full condition. For example: ICD code 01166 represents TB pneumonia. Which means fluid collection in lungs due to TB. however, ICD code 01170 represents TB pneumothorax which means damage of lungs due to TB. Any ICD 9 code starting with 011 represent conditions happening due to TB. It is enough for the therapist to know that the patient has TB to start the treatment. ADMITTIME is subtracted from DISCHTIME in order to get LOS (Length of Stay) and then both these columns are dropped. All the rows where ISREEOR is true are dropped A new column called mortality is created which is it is marked as "died" if DEATHTIME is not NaN else marked as "discharged". Only the rows which are marked as marked as "Discharge summary" in CATEGORY are retained and rest of them are dropped.

C. *Word2VEC*

The cleaning of the text column is done with the help of word2vec algorithm. Initially, to avoid the conflict with cases, the entire text is converted to lower case. Further, in order to avoid the empty words, when there are many whitespaces, they are being replaced with a single whitespace. Fig.2 highlights this process where the discharge summary is considered as an input followed by data cleaning and further Deep Neural Network (DNN) Autoencoder is used for exploring Term Frequency (TF) and Inverse document frequency (IDF). This process yield cluster identity which is further used for performing classification in next step.

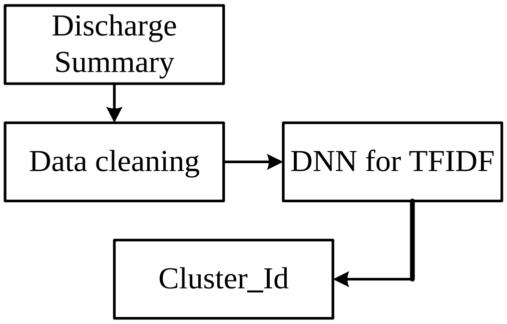

Fig. 2. Block Diagram of the Word2vec Algorithm used in Proposed Work.

As a standard procedure in the NLP, the punctuation marks are removed however the stop words are kept intentionally. Stop words are the words which won't carry lots of meaning ex: in, it, is and etc. They are not removed since they are important while recognizing the person's level of knowledge of medical science. In most of the cases, the patient testimony contains lots of non-technical words whereas clinical notes contain more specific technical words. Once these steps are done, the data is now in a format to be processed by NLP algorithms. Hence, once it is ready, TF-IDF vectorization is applied in the standard method. DNN is trained in such a way that it clusters the data based on the language used. DNN is used to recognize text if it is a patient testimony or clinical notes. The DNN here is setup here in autoencoder configuration. Cluster id 0 represent patient testimony and cluster id 1 represent clinical notes. Here W2V is being trained purely based on the style of writing used in this study: Autoencoder. Autoencoder is a type of DNN which gives exact same output as input. Autoencoder is generally used for data compression. Generally, there are odd numbers of hidden layers in an autoencoder and in this case, there are 5 hidden layers. General applications of AE are, I) Data compression ii) Data de-noising iii) Data generation iv) Data clustering. Data clustering is relatively new technique but not a novel technique. In present study, AE is configured to perform the same.AE is trained with only one class of data. (Either clinical notes or Patient testimony) In this case, AE is trained with Clinical notes. The Clinical notes is text and so is the Patient testimony. When the AE is trained with clinical notes only, then when it is given a Patient testimony as input, it will try to represent that text as a Clinical note. If the input is clinical note, then the output will be same as input However, if input is patient testimony, then output will be totally different compared to input. This is due the fact that AE is trained with only clinical notes. When input is clinical notes, the output is pretty similar. In other words, Euclidian distance between the input and output is less. When input is Patient testimony, output is different compared to input and Euclidian distance is more. We cluster the data based on this Euclidian distance. The threshold for Euclidian distance is set to 0.1. This is done by considering the least loss of the AE.

D. Design of Dual LSTM

This is the most important part of the study where the text from the discharge summary is classified into various diagnosis. This LSTM network contains an attention layer which changes its weights and biases as the admission type varies. The first set of weights and biases are used when the patient is admitted through OPD and second set is used when the patient is admitted through ER. The LSTM gives more importance to patient testimony when the patient is OPD. Depending on the cost center the attention layer allows either first set of weights and biases or 2nd set of weights and biases to process the data. Hence, we will have two specialized neural networks in one. Hence this particular network becomes more robust while diagnosing the patient based on description. Figure 3 shows the architecture of proposed dual LSTM which combines work encoding layer, attention layer, two LSTM layer, one CNN layer and single output layer.

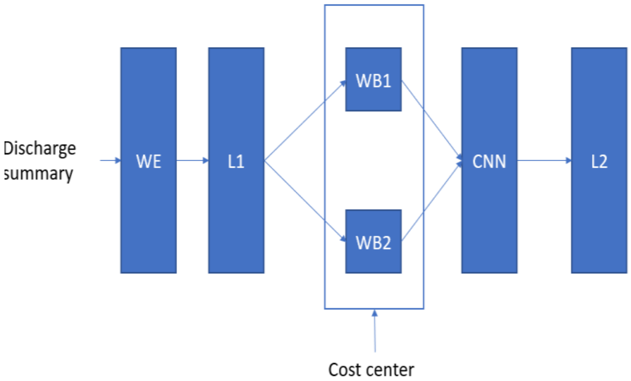

Fig. 3. The Novel Dual LSTM for Processing Medical Data.

V. Result Analysis

The proposed system's design and development is carried out using Python programming language and an anaconda computing environment. This section discusses the outcome and performance analysis of the proposed system.

The above figure 4 shows the analysis regarding number of words per sample versus percentage of sample. IThe graph trend exhibits that the most clinical notes contain more than 1,000 words. This is unsurprising as the documents are discharge summary and they contain all details from admission to discharge.

Figure 5 shows an analysis of the number of words versus a number of documents. The graph trend exhibits the number of documents in which so many words are present and there are rare documents containing 8000 words. However, it can be seen that most documents contain 1000 words.

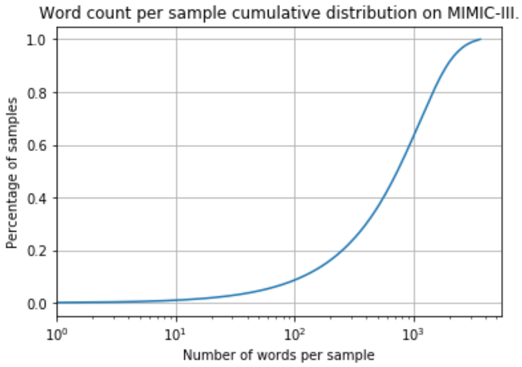

Fig. 4. Word Count per Sample.

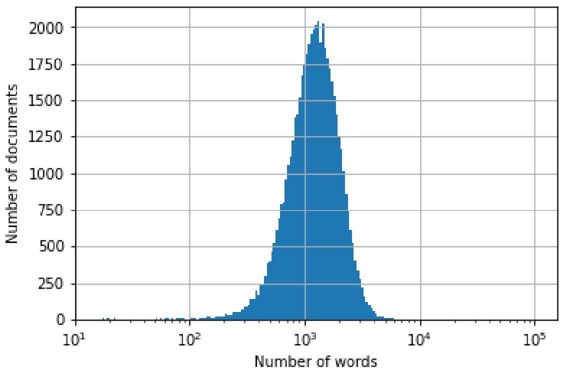

Fig. 5. Number of Words Histogram.

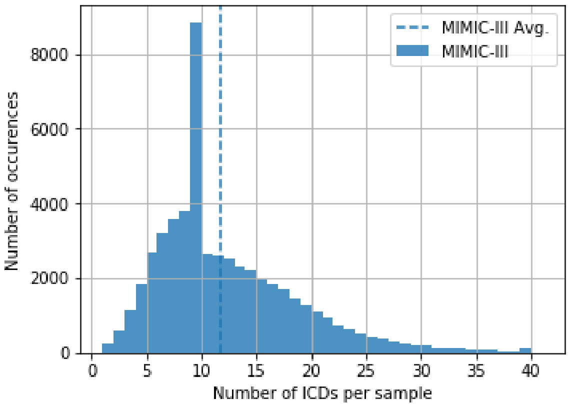

Fig. 6. ICD Histogram.

Figure 6 shows an analysis of the probability distribution of the ICD codes. Based on the graphical trend and frequency of occurrences, data is highly imbalanced, and some ICD codes have been repeated to 8,000 times, whereas others are repeated a very handful number of times. The proposed dual LSTM algorithm handles this imbalance.

The figure 7 shows the analysis regarding the number of classes from most to least occurrences versus frequency. The graph trend exhibits that the less than 100 classes have the highest frequency of occurrences. At the same time, most of the classes have fewer and more frequent occurrences.

The analysis from figure 8 shows that the proposed method shows a better result than all other existing methods. This is due to the fact that we are using LSTM, and text data is a series data. LSTM works best for the series data. The proposed algorithm shows a better result than the other shallow learning and deep learning methods with improvements. As it can be observed, the recall rate has improved greatly. This is because in this study, the more precise and technical clinical notes are given preference; hence, the number of false negatives reduces greatly. As the complete work is carried out on standard MIMIM-III dataset, which is universally approved, the applicability of the proposed scheme suits well with all kind of real-time dataset, which is structured in the form of MIMIC-III dataset or with slightest amendment. It can be used for diagnosis of any form of critical disease using ICD9 codes.

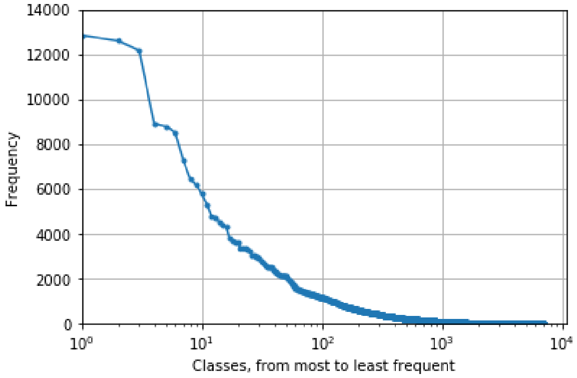

Fig. 7. Classes from Most to Least Frequent.

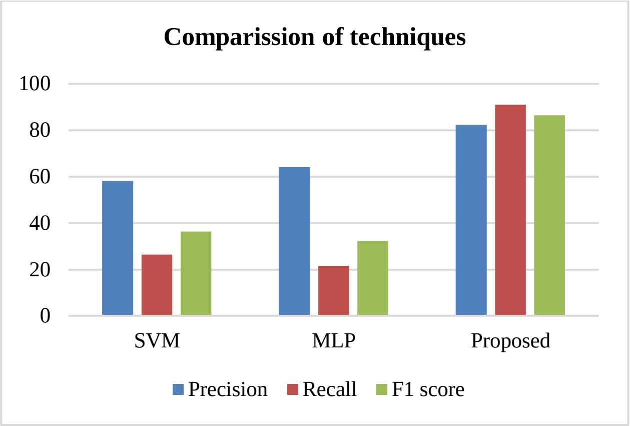

Fig. 8. Compression of Various Techniques.

VI. Discussion

The previous section has exhibited the outcome (both individual and comparative) which states proposed scheme excels better performance with respect to existing learning scheme. However, some important learning outcomes are obtained from this which is stated as follows:

- The study considers the performance parameters of precision, recall, and F1-score instead of choosing convention accuracy parameter owing to potential imbalance in classification using multiple clinical attributes.
- Inspite of potential capability to process unstructured data by SVM, still its performance is degraded as it offers challenges in fine tunning hyper parameters. This causes declination of precision (Table II). However, owing to its capability to overcome over-fitting data, its recall rate is better compared to MLP technique.
- Higher Precision score for MLP attributes for its capability to offer robustness against error prone data as well as due to presence of target function, it can offer discrete outcome (Table II). However, MLP is better suited for numerical data whereas the data considered in proposed scheme has both strings and numerical which reduces the recall rate as well as F1-Score.

- Proposed system offers a progressive scheme which uses autoencoders as well as CNN for learning purpose without much reengineering process involved in feature management. This potentially results in improved performance in every aspect (Table II).

TABLE II. COMPARATIVE ANALYSIS

Algorithms	Precision	Recall	F1 score
SVM	58.21	26.36	36.28747
MLP	64.03	21.58	32.28051
Proposed	82.32	90.96	86.4246

Therefore, usage of dual LSTM significantly assisted in overcoming the diagnosis prediction problem that is explored in existing review of literature. However, the primary challenge encountered in the proposed study was to carry out preparation of the data prior to subjecting it to learning operation. This challenge is mitigated by its first module of data preparation where a selected field from the dataset is considered followed by using deep neural network autoencoder.

VII. CONCLUSION

This paper has presented an effective learning system to support the clinical decision process in the patient diagnosis. The proposed system is advanced and highly optimized to process the clinical notes written in rich language. The contribution made in this paper are as follows: i) suitable data treatment and cleaning operation is applied to clinical notes for the processing of NLP and learning mechanisms; ii) Work2vec modeling with Autoencoder is applied to perform clustering of the two distinct clinical classes for the predictive modeling and iii) a dual LSTM is built based on the joint operation of LSTM and CNN deep learning approach. The study outcome exhibited higher performance achieved by the proposed system compared to the shallow machine learning approaches.

REFERENCES

[1] Bhasale A. The wrong diagnosis: identifying causes of potentially adverse events in general practice using incident monitoring. Family Practice. 1998 Aug 1;15(4):308-18.

[2] Rogers WA. Is there a moral duty for doctors to trust patients? Journal of Medical Ethics. 2002 Apr 1;28(2):77-80.

[3] Day SC, Cook EF, Funkenstein H, Goldman L. Evaluation and outcome of emergency room patients with transient loss of consciousness. The American journal of medicine. 1982 Jul 1;73(1):15-23.

[4] Applebaum GE, King JE, Finucane TE. The outcome of CPR initiated in nursing homes. Journal of the American Geriatrics Society. 1990 Mar;38(3):197-200.

[5] Klein MH, Benjamin LS, Rosenfeld R, Treece C, Husted J, Greist JH. The Wisconsin personality disorders inventory: Development, reliability, and validity. Journal of Personality Disorders. 1993 Dec;7(4):285-303.

[6] Putra FB, Yusuf AA, Yulianus H, Pratama YP, Humairra DS, Erifani U, Basuki DK, Sukaridhoto S, Budiarti RP. Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11). In2019 International Electronics Symposium (IES) 2019 Sep 27 (pp. 1-5). IEEE.

[7] Waheeb SA, Ahmed Khan N, Chen B, Shang X. Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. Information. 2020 May;11(5):281.

[8] Diallo B, Hu J, Li T, Khan GA, Liang X, Zhao Y. Deep embedding clustering based on contractive Autoencoder. Neurocomputing. 2021 Apr 14;433:96-107.

[9] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

[10] https://people.cs.pitt.edu/~jlee/papers/cp1_survey_jlee_amuis.pd

[11] Mustafa, Akram, and Mostafa Rahimi Azghadi. "Automated Machine Learning for Healthcare and Clinical Notes Analysis." Computers 10, no. 2 (2021): 24.

[12] Kaur, R. A comparative analysis of selected set of natural language processing (NLP) and machine learning (ML) algorithms for clinical coding using clinical classification standards. Stud. Health Technol. Inform. 2018, 252, 73–79.

[13] Hassler, A., Menasalvas, E., García-García, F. et al. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak 19, 33 (2019). https://doi.org/10.1186/s12911-019-0747-6.

[14] Kashina, M., Lenivtceva, I.D. and Kopanitsa, G.D., 2020. Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification. Procedia Computer Science, 178, pp.284-290.

[15] Ferrão, José & Oliveira, Mónica & Janela, Filipe & Martins, Henrique. (2016). Preprocessing structured clinical data for predictive modeling and decision support: A roadmap to tackle the challenges. Applied Clinical Informatics. 7. 1135-1153. 10.4338/ACI-2016-03-SOA-0035.

[16] Misra, Puneet & Yadav, Arun. (2019). Impact of Preprocessing Methods on Healthcare Predictions. SSRN Electronic Journal. 10.2139/ssrn.3349586.

[17] Teo, Kareen & Yong, Ching & Chuah, Joon Huang & Murphy, Belinda & lai, khin wee. (2020). Discovering the Predictive Value of Clinical Notes: Machine Learning Analysis with Text Representation. Journal of Medical Imaging and Health Informatics. 10. 2869-2875. 10.1166/jmihi.2020.3291.

[18] Spasic, Irena, and Goran Nenadic. "Clinical text data in machine learning: systematic review." JMIR medical informatics 8, no. 3 (2020): e17984.

[19] Ye, J., Yao, L., Shen, J. et al. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak 20, 295 (2020). https://doi.org/10.1186/s12911-020-01318-4.

[20] Weng, WH., Wagholikar, K.B., McCray, A.T. et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 17, 155 (2017). https://doi.org/10.1186/s12911-017-0556-8.

[21] V. Kumar, D. R. Recupero, D. Riboni and R. Helaoui, "Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes," in IEEE Access, vol. 9, pp. 7107-7126, 2021, doi: 10.1109/ACCESS.2020.3043221.

[22] Topaz, Maxim, et al. "Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches." Journal of biomedical informatics 90 (2019): 103103.

[23] Poulin, Chris, et al. "Predicting the risk of suicide by analyzing the text of clinical notes." PloS one 9.1 (2014): e85733.

[24] Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." arXiv preprint arXiv:1904.05342 (2019).

[25] Prabhakar, S. K., & Won, D. O. (2021). Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention. Computational Intelligence and Neuroscience, 2021.

[26] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar and G. Srivastava, "An Accurate Deep Learning Model for Clinical Entity Recognition From Clinical Notes," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 10, pp. 3804-3811, Oct. 2021, doi: 10.1109/JBHI.2021.3099755.

[27] Xu, K.; Lam, M.; Pang, J.; Gao, X.; Band, C.; Mathur, P.; Papay, F.; Khanna, A.K.; Cywinski, J.B.; Maheshwari, K. Multimodal machine

learning for automated ICD coding. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Ann Arbor, MI, USA, 8–10 August 2019; pp. 197–215.

[28] Soguero-Ruiz, C.; Hindberg, K.; Rojo-Álvarez, J.L.; Skrøvseth, S.O.; Godtliebsen, F.; Mortensen, K.; Revhaug, A.; Lindsetmo, R.O.; Augestad, K.M.; Jenssen, R. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. IEEE J. Biomed. Health Inform. 2014, 20, 1404–1415.

[29] Yogarajan, V.; Montiel, J.; Smith, T.; Pfahringer, B. Seeing The Whole Patient: Using Multi-Label Medical Text Classification Techniques to Enhance Predictions of Medical Codes. arXiv 2020, arXiv:2004.00430.

[30] Karmakar, A. Classifying medical notes into standard disease codes using Machine Learning. arXiv 2018, arXiv:1802.00382

[31] Wei, Q.; Ji, Z.; Li, Z.; Du, J.; Wang, J.; Xu, J.; Xiang, Y.; Tiryaki, F.; Wu, S.; Zhang, Y. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J. Am. Med. Inform. Assoc. 2020, 27, 13–21.

[32] Polignano, M.; Suriano, V.; Lops, P.; de Gemmis, M.; Semeraro, G. A study of Machine Learning models for Clinical Coding of Medical Reports at CodiEsp 2020. In Proceedings of the Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, Thessaloniki, Greece, 2–25 September 2020.

[33] Lin, C.; Hsu, C.J.; Lou, Y.S.; Yeh, S.J.; Lee, C.C.; Su, S.L.; Chen, H.C. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. J. Med. Internet Res. 2017, 19, e380.

[34] Garla, V.N.; Brandt, C. Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. 2012, 45, 992–998.

[35] Cobb, R.; Puri, S.; Wang, D.Z.; Baslanti, T.; Bihorac, A. Knowledge extraction and outcome prediction using medical notes. In Proceedings of the ICML Workshop on Role of Machine Learning in Transforming Healthcare, Atlanta, GA, USA, 20–21 June 2013.

[36] Shi, H.; Xie, P.; Hu, Z.; Zhang, M.; Xing, E.P. Towards automated ICD coding using deep learning. arXiv 2017, arXiv:1711.04075.