Improving Clinical Efficiency and Reducing Medical Errors through NLP-enabled Diagnosis of Health Conditions from Transcription Reports

Krish Maniar[1], Shafin Haque[2], Kabir Ramzan[3]

[1,3]The Harker School in Saratoga, California, United States

[2]Saratoga High School in Saratoga, California, United States

ABSTRACT

Misdiagnosis rates are one of the leading causes of medical errors in hospitals, affecting over 12 million adults across the US. To address the high rate of misdiagnosis, this study utilizes 4 NLP-based algorithms to determine the appropriate health condition based on an unstructured transcription report. From the Logistic Regression, Random Forest, LSTM, and CNN-LSTM models, the CNN-LSTM model performed the best with an accuracy of 97.89%. We packaged this model into a authenticated web platform for accessible assistance to clinicians. Overall, by standardizing health care diagnosis and structuring transcription reports, our NLP platform drastically improves the clinical efficiency and accuracy of hospitals worldwide.

Index Terms : Natural Language Processing, Medical Tran-scription Notes(s), Diagnostic Systems, Recurrent Neural Net-work, Long Short-Term Memory Network

I. INTRODUCTION

A. Physician Ineffectiveness in Hospitals

In the United States alone, nearly 400,000 people are killed annually due to preventable medical errors [1]. With over 12 million adults in the U.S. receiving misdiagnoses yearly, inaccurate diagnoses are the leading cause of medical errors in hospitals [2]. Misdiagnoses refer to inaccurate assessments by healthcare providers of a patient's condition, often leading to either inappropriate or excessive treatment and sometimes no treatment at all. In a case study conducted at the University of Maryland Medical Center, it was found that in a group of 177 patients, over 90% of patients received at least one unnecessary treatment, highlighting the unreliable nature of physicians in hospital settings [3]. Not only can misdiagnoses undermine the effectiveness of clinicians, but over one-third of such cases with inaccurate treatment result in death, injury, or permanent disabilities [4].

Beyond high misdiagnosis rates, the lack of communication between patients and doctors contributes to diminished patient outcomes and care. In 2010, a study conducted in JAMA Internal Medicine discovered the striking discrepancies in communication between patients and doctors, especially in relation to diagnoses. The authors determined that while 77% of physicians believed patients knew their diagnosis, only 57% of patients

actually did [5]. This alarming inconsistency reveals broader issues regarding the subjectivity, unreliability, and variability between clinicians, demanding the necessity for more objective and accurate diagnostic approaches.

B. Electronic Health Records and Unstructured Data

The recent rise in electronic health records (EHRs) only increases this variability and further warrants the need for automated methods to standardize diagnosis procedures in medical settings [6]. EHRs primarily consist of unstructured data, which lacks standardized specifications and is written in free-text without, clear separations. Unstructured data is often found in clinical report notes, discharge summaries, and chart reviews; however, interpreting such data requires costly time and presents a considerable challenge to medical practitioners due to its lack of baselines and standardized components. In fact, the The Agency for Healthcare Research and Quality acknowledges unstructured data in EHRs as a primary barrier in understanding data quality and interpreting key diagnostic information [7]. Currently, most teams and researchers that focus on medical record abstraction target structured data, ignoring the widespread amount of unstructured data.

Because of the sheer amount of unstructured data available, the use of Machine Learning and Deep Learning methods are crucial for eliminating clinician subjectivity across hospitals around the world and for utilizing the vast amount of data available to improve diagnosis.

II. LITERATURE REVIEW

A. Emergence of Machine Learning and NLP

Machine Learning, a subset of Artificial Intelligence, refers to algorithms that can efficiently learn, generalize, and thus develop a complex, mathematical understanding of pre-existing training data to give accurate predictions on unseen test data

[8]. As a result of the increase in computational power and the emergence of the "big data" era, the application of ML in medical imaging and health care systems has increased and can strongly facilitate healthcare operations [9]. More specifically, Natural Language Processing (NLP), a form of Deep Learning, has emerged as a promising discipline to analyze unstructured data, a task that is time-consuming and extremely challenging for doctors and physicians, especially in low-income countries.

Over 80% of all healthcare data is unstructured [10], so the ability of NLP models to decode unstructured clinical reports and provide insightful diagnostic advice presents an enormous opportunity. However, NLP is most often associated with speech analytics, sentiment analysis, and chat generation, with relatively little attention for misdiagnosis rates in medical settings. As the healthcare community looks to incorporate exponentially more automated technologies into their infras-tructure, it is vital to approach specific problems in the medical community with state-of-the-art NLP models.

B. Objective

While recent studies have used models such as RNNs [11] to predict surgery complications and extract structured data from clinical reports, very few institutions have focused their research on correcting misdiagnosis rates. Moreover, there is a lack of an easily accessible and accurate early-diagnostics system for classifying medical specialties from unstructured medical transcription notes.

Thus, in an effort to streamline the process for medical practitioners and improve clinical care for patients within the hospital, we developed various machine learning and deep learning-based architectures to determine the specific medical specialty that a clinical transcription note describes, ultimately reducing misdiagnosis rates and alleviating the burden for overloaded hospitals. Furthermore, to increase accessibility for hospitals worldwide, our

models are packaged into an efficient and accurate AI-enabled medical app.

III. METHODS

To categorize the medical transcription reports into various specialties and types, our machine learning models require large amounts of categorical data to operate smoothly. In addition to scraping data from an online website to curate our dataset, we also implemented various types of data pre-processing to maximize the performance of all models used.

A. Data Collection and Transformation

We retrieved and scraped our data from MTSamples, a public corpus of data that contains transcribed medical reports. 5003 sample reports were extracted from the MTSamples collection of data. It contains 40 different medical specialties, such as 'Urology' and 'Nephrology,' 28581 unique words, and 487.5 average characters per sentence.

To reduce heterogeneity between values in the dataset, we funneled the data down to four main human body systems: Heart, Brain, Reproductive, and Digestive. To maintain a fair balance of nearly 300 reports per category, we accordingly combined certain medical specialties. For example, the Brain label contained both "Neurology" and "Neurosurgery" reports, and the Digestive label contained both "Gastroenterology" and "Nephrology." In total, our modified dataset had 1304 reports (371 heart, 317 brain, 311 reproductive, and 305 digestive), a unique vocabulary size of 20127, and a mean sentence length of 464.9 characters. A sentence length histogram is displayed in Figure 1 to illustrate the distribution of sentence length for the medical transcription reports included in our modified dataset.
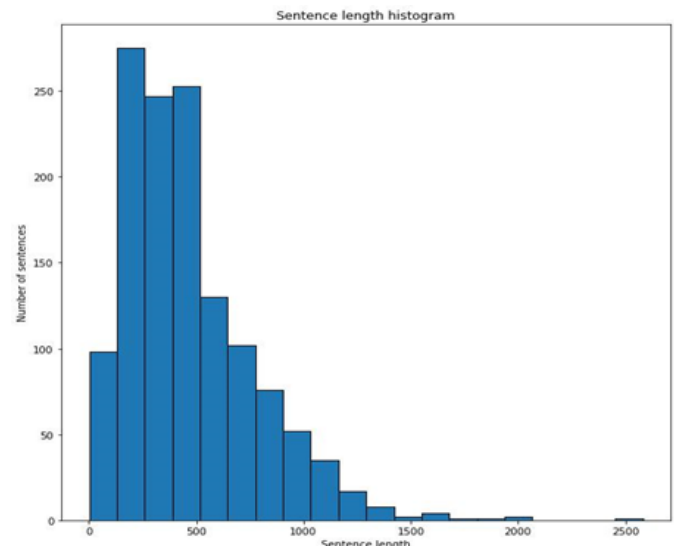

Fig. 1. Transcription Report Sentence Length Histogram

Within our dataset, each report contains a variety of non-standardized labels with sections differing even within med-ical specialties. Our scraper attempted to extract all relevant information regarding the medical specialty and transcription report details. To further resolve any discrepancies and correct errors from our scraping algorithm, we also screened the full modified dataset.

B. Data Preprocessing

After transforming the dataset and parsing the specified categories, we applied four preprocessing steps on cleaning the data. First, we removed any reports with missing transcription details or medical specialties. Second, we removed numbers, symbols, punctuation, and special characters such as brackets and slashes to standardize the text and delete unnecessary elements, allowing for improved analysis of the transcription input. Third, we tokenized our transcription sentences into smaller tokens, all in lower-case, while also removing stop-words that lack contextual significance. Fourth, we lemmatized our data by converting the tokenized elements into shorter but more useful forms and linking words to its root through a dictionary-based implementation. These four steps work to-gether to convert the

initially unstructured transcription report to a more standardized and structured report.

Once our data was cleaned, we applied a recently developed vectorization preprocessing algorithm called Term Frequency Inverse Document Frequency (TFIDF) to translate the string tokens to numerical vectors. Most natural language processing models use the Bag Of Words (BOW) algorithm, where the location of words directly influences the text vectorization. However, this approach fails to utilize semantics to guide the vectorization of an embedding space. The TFIDF algorithm builds upon this traditional BOW approach and consists of two primary steps: Term Frequency and Inverse Data Frequence. The first step divides the frequency of each word by the total number of words in the text. The second step weights each word based on its semantic relevance to the medical specialty. For example, words like "the" and "of" may be the most frequent tokens but lack semantic value to the classification task, so they will be assigned a low weight. Both steps work together to produce a vector and represent the reports in an embedding space. Figure 1 displays the embedding space for all four categories in the dataset (Brain, Heart, Digestive, and Reproductive).
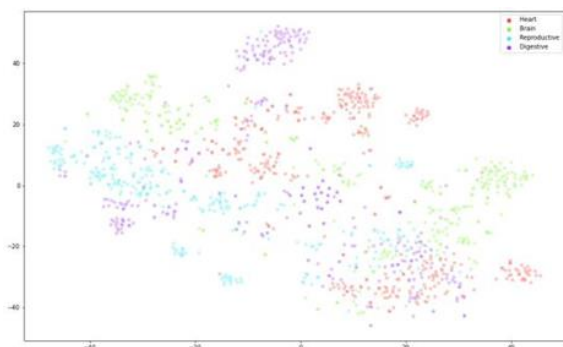

Fig. 2. TFIDF Vector Embeddings for Modified Dataset

While the embeddings are segregated in certain portions of the figure, they overlap in numerous areas, which both explains the discrepancies that doctors have in classifying such reports and adds to the complexity of our data and models.

C. Machine Learning Models

Once we processed our data into an appropriate format, we developed 2 ML and 2 DL models. All models were trained and tested with a dataset consisting of an 80% train and a 20% test split.

Logistic Regression. Our first ML algorithm (Logistic Regression) generally uses a logistic function to separate two classes from each other through using a binary entropy function. However, our dataset contains multiple classes, so we modified our loss function from binary to categorical class entropy, a multinomial probability distribution function that computes the loss over all four classes. The loss function formula is displayed in Figure 3.

$$\text{Loss} = -\frac{1}{\substack{\text{output} \\ \text{size}}} \sum_{i=1}^{\substack{\text{output} \\ \text{size}}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

Fig. 3. Category Cross Entropy Loss Function

We also applied Principal Component Analysis with 95% variance to reduce the dimensionality of the features being inputted into the model. This helped expedite the training process and improve the interpretability of our Logistic Re-gression model.

Random Forest. In addition to Logistic Regression, we used Random Forest, another common Machine Learning architecture. As shown in Figure 4, this model consists of an ensemble of decision trees and uses numerous yes/no branches to deliver a result for classification problems. After fine-tuning several hyper parameters, namely the number of decision trees and respective branch depth, we determined the optimal amount of estimators was 150 trees and ideal maximum amount of depth was 4 branches.
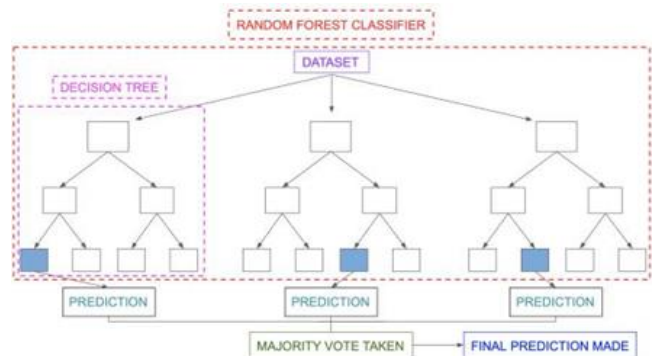

Fig. 4. Random Forest Architecture [12]

D. Deep Learning Models

ML models such as Logistic Regression lack the ability to interpret semantics and maintain long-term connections in text. Thus, we also developed several Deep Learning methods: an LSTM and a CNN-LSTM, which combine an LSTM's semantic elements and a CNN's spatial dependencies.

Long Short-Term Memory Network. By constructing long-term connections in the transcription reports, the LSTM model is an improvement over Recurrent Neural Networks. A key difference between LSTM and RNN architectures is the presence of a memory cell that maintains a hidden state as it progresses through the text. Figure 5 illustrates the details of a memory cell. The states are regulated by the forget, input gate, and output gate, which alter memory connections between segments of the data.

CNN-LSTM. While CNN models are generally used for computer vision tasks, they can provide useful information from text if used in conjunction with an LSTM model. Because CNNs use convolutional layers and padding, they can extract higher-level features than LSTMs. However, they are unable to track long-term dependencies in transcription reports. Thus, a combination of CNN and LSTM models would be expected to outperform a core RNN or LSTM model. Similar training pa-rameters were utilized as the LSTM network, and an overview of the model architecture is visible in Figure 6.

E. MedicAI Web Application

To package our Machine and Deep Learning models into an accessible platform, we created a Flask web application called MedicAI that serves as a dashboard for clinicians. As shown in the dashboard in Figure 7, doctors are able to access medical transcript records for all their patients, and they can accordingly add and filter patient records. Once a record is uploaded as an image, our Optical Character Recognition (OCR) algorithm will tokenize the text from the image and parse the corresponding" Findings" section of the report. Then, as described in the 'Data Preprocessing' section, this parsed text will be preprocessed using TFIDF and subsequently inputted to the model for diagnosis classification.
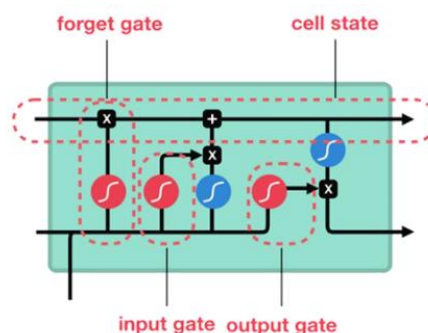

Fig. 5. LSTM memory cell and corresponding gates [13]
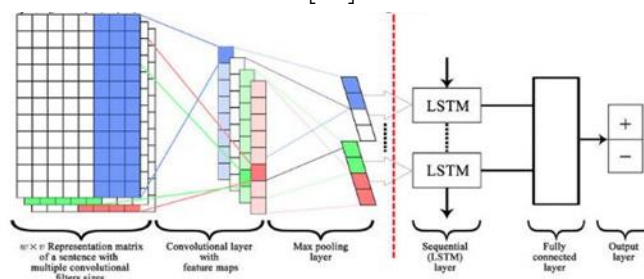

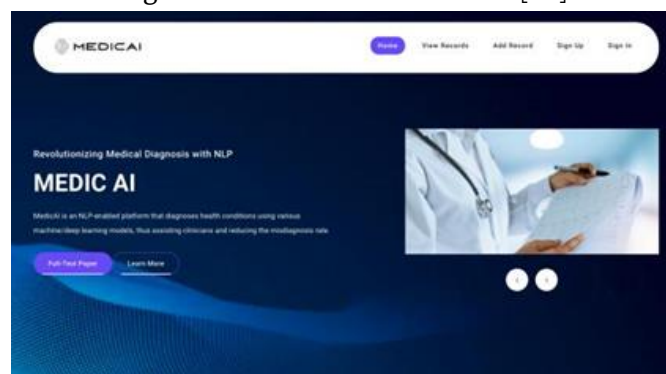Fig. 6. CNN-LSTM architecture [14]

Fig. 7. Dashboard for MedicAI Web Application

In terms of the web technologies we used, our front-end was made with HTML, CSS, and JavaScript, and the back-end infrastructure was developed with Flask and Firebase. We im-plemented authentication procedures with secure login/logout functionality using Firebase and developed the database using Firebase Cloud Firestore. We believe our application is crucial for assisting clinicians in health condition diagnoses, and it has the potential of leveling the playing field between hospitals of different areas.

IV. RESULTS

Once we developed Logistic Regression, Random Forest, LSTM, and CNN-LSTM, we trained the models on 80% of the dataset and tested them on the remaining 20 percent. The Machine Learning models were developed with SciKit-Learn, and the Deep Learning models were developed with TensorFlow. All models were trained with a categorical cross-entropy loss, a softmax activation function, and 50 epochs. To analyze the performance of each model, we calculated 4 commonly used performance metrics: Accuracy, Precision, Recall, and F1-score. Figure 8 displays a detailed breakdown of the formulas for each performance metric based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rates.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Fig. 8. Formulas for Performance Metrics

After tuning the Logistic Regression model's hyperparame-ters with a gridsearch function, the model yielded an accuracy of 91.19% accuracy. Table I describes the results for the other performance metrics.

TABLE I LOGISTIC REGRESSION PERFORMANCE ON THE TEST DATASET

	Precision	Recall	F1-score
Heart	0.92	0.95	0.93
Brain	0.94	0.94	0.94
Reproductive	0.90	0.89	0.89
Digestive	0.88	0.87	0.88

As for the Random Forest model, the resulting accuracy on the test set was 84.29% accuracy. Other performance metrics are displayed in Table II. The optimal hyperparameters were 150 decision trees, each with a depth of 4 branches. The other performance metrics were also computed in Table II.

For most of the metrics, the Logistic Regression and Ran-dom Forest models performed the best for the Brain class, with lower performance on the Reproductive and Digestive classes. Additionally, the more traditional Logistic Regression performed with an accuracy of nearly 7% higher, suggesting its simplicity outperforms the complex ensembled decision tree estimators.

TABLE II RANDOM FOREST PERFORMANCE ON THE TEST DATASET

	Precision	Recall	F1-score
Heart	0.81	0.91	0.85
Brain	0.89	0.92	0.91
Reproductive	0.86	0.77	0.79
Digestive	0.82	0.75	0.79

While the two Machine Learning models performed at an acceptable accuracy, the Deep Learning methods——LSTM and CNN-LSTM——produced much higher performance rates. The Deep Learning models were both trained for 50 epochs on a 20% train set with a batch size of 100. The LSTM model performed at an accuracy of 97.60%, and the CNN-LSTM performed slightly better with an accuracy of 97.89%. More detailed performance metrics are illustrated in Table III and Table IV below for the LSTM and CNN-LSTM model, respectively.

TABLE III LSTM PERFORMANCE ON THE TEST DATASET

	Precision	Recall	F1-score
Heart	1.00	0.97	0.98
Brain	0.98	1.00	0.99
Reproductive	0.97	0.98	0.97
Digestive	0.96	0.96	0.96

TABLE IV CNN-LSTM PERFORMANCE ON THE
TEST DATASET

	Precision	Recall	F1-score
Heart	1.00	0.97	0.98
Brain	0.99	1.00	0.99
Reproductive	1.00	0.96	0.98
Digestive	0.93	0.99	0.96

Overall, the CNN-LSTM performed the best for most performance metrics, even outperforming the standard LSTM model. This is likely due to the spatial features of the CNN that facilitate the semantical and long-term memory connections computed by the LSTM model.

V. DISCUSSION

In this study, we developed a web application for doctors that utilizes both traditional and modern Machine Learning methods to improve diagnosis of health conditions from medical transcription notes. Not only does this app reduce misdiagnosis rates and increase clinical efficiency in hospitals, but it also standardizes medical diagnosis across hospitals around the world with an affordable and accurate tool.

After constructing a Logistic Regression, Random Forest, LSTM, and CNN-LSTM model, we determined that the CNN-LSTM model, which incorporates spatial features and long-term memory connections, performed the best with an accu-racy of 97.89% and an average F1-score of 0.98. Our model was able to seamlessly correctly classify Brain, Heart, Repro-ductive, and Digestive categories. We integrated this CNN-LSTM model into our platform called MedicAI to classify health conditions in real-time based on medical transcription notes, satisfying the goals of this project.

Further research into state-of-the-art NLP models and pre-processing methods should be explored. In the future, we should look into modern NLP models, such as BERT and GPT, which utilize transformer architectures and self-supervised learning to learn improved representations. To support the computational complexity brought upon by these models, we also hope to develop our back-end infrastructure to support larger ML models. Additionally, we plan to test varying preprocessing methods for converting tokens into vectors in an embedding space, such as Glove and Word2Vec. Lastly, we hope to expand the available data and scope of the project by incorporating institutional data from hospitals around the world, and eventually implementing this platform in global medical centers. Thus, we hope this project serves as a strong basis for future research into the use of advanced NLP methods for medical specialty classification from transcription notes.

VI. REFERENCES

[1]. J. T. James, "A new, evidence-based estimate of patient harms associated with hospital care," J. Patient Saf., vol. 9, no. 3, pp. 122–128, Sep. 2013, doi: 10.1097/PTS.0b013e3182948a69.

[2]. H. Singh, A. N. D. Meyer, and E. J. Thomas, "The frequency of diagnos-tic errors in outpatient care: estimations from three large observational studies involving US adult populations," BMJ Qual. Saf., vol. 23, no. 9, pp. 727–731, Sep. 2014, doi: 10.1136/bmjqs-2013-002627.

[3]. C. Koch, K. Roberts, C. Petruccelli, and D. J. Morgan, "The Frequency of Unnecessary Testing in Hospitalized Patients," Am. J. Med., vol. 131, no. 5, pp. 500–503, May 2018, doi: 10.1016/j.amjmed.2017.11.025.

[4]. Johns Hopkins Medicine Researchers Identify Health Conditions Likely to be Misdiagnosed. 2019. [Online]. Available: https://www.hopkinsmedicine.org/news/newsroom/news-releases/johns-hopkins-medicine-researchers-identify-health-conditions-likely-to-be-misdiagnosed

[5]. D. P. Olson and D. M. Windish, "Communication discrepancies between physicians and hospitalized patients," Arch. Intern. Med., vol. 170, no. 15, pp. 1302–1307, Aug. 2010, doi: 10.1001/archinternmed.2010.239.

[6]. J. Howard et al., "Electronic health record impact on work burden in small, unaffiliated, community-based primary care practices," J. Gen. In-tern. Med., vol. 28, no. 1, pp. 107–113, Jan. 2013, doi: 10.1007/s11606-012-2192-4.

[7]. http://www.ahrq.gov/qual/prospectscare/prospects1.htm.

[8]. R. C. Deo, "Machine Learning in Medicine," Circulation, vol. 132, no. 20, p. 1920, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[9]. D. Shen, G. Wu, and H.-I. Suk, "Deep Learning in Medical Image Analysis," Annu. Rev. Biomed. Eng., vol. 19, p. 221, Jun. 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[10]. L. Rasmy et al., "Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data," arXiv, Oct. 2021, doi: 10.13140/RG.2.2.13199.51368.

[11]. H.-J. Kong, "Managing Unstructured Big Data in Healthcare System," Healthcare Informatics Research, vol. 25, no. 1, p. 1, Jan. 2019, doi: 10.4258/hir.2019.25.1.1.

[12]. Introduction to Random Forest in Machine Learning. 2022. [Online]. Available: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning

[13]. Long Short-Term Memory (LSTM). 2022. [Online]. Available: http://www.bhlhdb.org/content/8/LSTM

[14]. A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A Combined CNN and LSTM Model for Arabic Sentiment Analysis," in Machine Learning and Knowledge Extraction, Cham, Switzerland: Springer, 2018, pp. 179–191. doi: 0.1007/978-3-319-99740-7 12.