* 1. **Introduction**

**A**ccurate classification of diseases is not merely a task of assigning labels but it is also the cornerstone of effective healthcare. As it lays a foundation for improving patient’s health by performing precise diagnosis and targeted treatment. As timely classification of symptoms is important for betterment of the patient .As nowadays it is challenging to determine the meaning of symptoms and to seek appropriate advice from doctors. Under this context NLP targets at extracting a useful information from the user’s text . NLP methods are used for the communication between computer and human language , as it involves transforming text into numeric data which is used in data mining algorithms. NLP techniques makes possible to extract valuable insights from patient narratives to make more informed decisions.

Classifiers utilize advanced algorithms and machine learning techniques to analyze user input and then classify it accordingly . It includes various benefits for both users as well as healthcare providers , for users it is easy for use ,convenient to identify symptom and to get personalize recommendation and relevant information about healthcare providers. Healthcare providers get benefits to streamline the process and to identify the patterns in the reported symptoms ,and also can help in prioritizing patients based on severity of symptoms. As early detection of potential health concern ,allows for preventive measures , and allowing healthcare professionals to diagnose at early stage , to reduce the impacts caused by it in future. Accurate classification also helps the patient as it provides clarity and reassurance ,which reduces the uncertainty.

Timely classification of symptoms allows for early detection of potential health concerns and accurate symptom classification enhances the efficiency of healthcare delivery by optimizing resource allocation and reducing unnecessary diagnostic procedures. By streamlining the diagnostic process, healthcare systems can allocate resources more effectively, ensuring that patients receive timely access to appropriate care.

The potential applications of symptom classifiers are vast and diverse, spanning various domains within healthcare. From assisting the healthcare providers in differential diagnosis and treatment planning to empowering patients with self-assessment tools and also for doctor recommendation, symptom classifiers hold the promise of revolutionizing the way symptoms are recognized, interpreted, and acted upon in clinical practice.

This paper will explore the development, applications, and implications of symptom classifiers in healthcare, with a specific focus on their potential to shape the future of accurately classifying symptoms. It will delve into the evolution of symptom classification technologies.

* 1. **Motivation**

The motivation behind developing a symptom classifier , is to enhance the diagnostic accuracy and to improve patient care . To solve the problem of delayed diagnosis which poses significant risks to patient‘s safety and well-being, and to optimize treatment decisions, ultimately leading to better outcomes for patients. To help doctors with identifying the symptoms , and make their part easy ,which can help in prioritizing the patient’s based on the severity of the disease.

* Improving Diagnostic Accuracy : One of the primary motivations is to enhance diagnostic accuracy in healthcare. Accurate symptom classification is crucial for identifying underlying medical conditions and guiding to appropriate treatment decisions. By using NLP techniques to analyze and interpret textual descriptions of symptoms, the project aims to develop a tool that can assist healthcare providers in making more accurate and timely diagnoses.
* Addressing Healthcare Challenges : The project is motivated by the need to address significant challenges in healthcare, such as misdiagnosis, delayed diagnosis. Misdiagnosis can lead to serious consequences for patients, including inappropriate treatments, prolonged suffering, and even mortality. By developing a symptom classifier that can reliably categorize symptoms and suggest potential medical conditions, the project aims to solve these challenges and improve patient outcomes.
* Enhancing Patient Care and Experience : Another key motivation is to enhance patient care and experience. Prompt and accurate diagnosis is essential for providing timely treatment and relieving patient distress. By providing healthcare providers with a tool that can assist in symptom classification, and ultimately improve the overall patient experience.
* Providing doctor recommendations : The aim is to provide doctor recommendations based on the symptoms identified. By integrating this feature, the project seeks to streamline the patient journey by connecting individuals with healthcare professionals who specialize in addressing their specific symptoms and medical conditions. This not only facilitates timely access to appropriate care but also enhances patient satisfaction and engagement in the healthcare process.
  1. **Objectives**
* Conduct exploratory data analysis on a dataset comprising over 6000 transcripts, focusing on understanding the dataset structure, identifying label inaccuracies, and addressing poor-quality transcripts.
* Prepare the dataset for training a natural language processing (NLP) model by preprocessing the data, including text tokenization ,feature extraction and implementing a proper train-validation-test split to optimize the accuracy and efficiency of symptom classification.
* Integrate the symptom classifier into healthcare systems to provide personalized symptom assessment and doctor recommendations to patients, thereby improving access to care and enhancing patient outcomes.
* Continuously refine and update the symptom classifier based on feedback from users, new data sources, and advancements in NLP technology to ensure its relevance and effectiveness in addressing evolving healthcare needs.
  1. **Problem Definition**

Developing an accurate system for the classification of medical symptoms from text inputs provided by the user’s and implementing NLP techniques and other machine learning algorithms. The objective is to create a model that can effectively analyze and categorize patient-reported symptoms, in written form ,and also to provide doctor recommendations to patients, thereby improving access to care and enhancing patient outcomes ,and to aid healthcare professionals in diagnosis and treatment planning.

* 1. **Brief Description of the System**

**Data Collection and Preprocessing :**

* **Data collection:** Gather relevant datasets containing symptom-related text . Explore medical literature, patient records, and publicly available healthcare databases and our source , <https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent> .
* **Data Preprocessing:** Clean and normalize the data , Tokenize text data.

**Feature Extraction :**

* **Text Features:**
  + TF-IDF (Term Frequency-Inverse Document Frequency): To represent text data.
  + Word Embeddings (Word2Vec, GloVe ): To capture semantic relationships between words.
  + BERT (Bidirectional Encoder Representations from Transformers): For contextualized embeddings.

**Model Selection and Training (Algorithm) :**

* **Text Classification Models :**
  + - Naive Bayes, SVM, or deep learning models (LSTM, BERT,Transformers).
* Split data into training, validation, and test sets.
* Train and fine-tune models using selected features.

**Evaluation Metrics :**

* **Define evaluation metrics :**
  + Accuracy, precision, recall, F1-score, AUC-ROC.
  + Consider the medical context and prioritize minimizing false negatives (missing symptoms).

**Model Evaluation and Validation :**

* Evaluate models on the validation set.
* Fine-tune hyperparameters.
* Validate models on a separate test set to assess generalization.

**User interface design :**

* Select a frontend framework or library (e.g., react) and set up the development environment.
* Design the user interface (UI) for data input (user interaction) and results display.
* Choose a backend framework (e.g., flask, django) and set up the development environment.

**2.1 Literature Survey**

Various papers had been studied related to Symptom Classifier about how they work. Some of the papers are mentioned below with their brief description of the studies.

**Paper 1**

**2022-**

**Identifying Symptom Information in Clinical Notes**

The paper tells about the current method for symptom extraction are labour intensive and lacks scalability , so a method combining vocabularies and NLP can address these challenges. It uses various approaches in the model like Nimble miner , word2vec ,phrase2vec ,random forest .It’s objective is to identify symptoms from EHR(Electronic Health Records ) and it has total 5 diverse symptoms . Formula used are as follows Performance  metrics  recall, precision, F-measure .The results of the nimble miner approach gives high recall , precision and f1-score , which significantly increases the accuracy for the classification of symptoms. The research gap in this paper is that the data is from a single medical center so less diversity , and the evaluation only on a subset of symptom ,where they have only five types of symptom .

**Paper 2**

**2021 -**

**Limitations of Transformers on Clinical Text Classification**

The paper tells about the limitations of transformers as they compare 2 much similar architectures , which are word level CNN and Hierarchial Self Attention Network (HSAN).

As each in word level CNN each word is transformed to a word vector with k-dimensions.

And in HSAN it performs word encoding , then word attention and same process for sentence .Formula used are weigth,bias vectors , softmax function. For Evaluation precision ,recall and F1 score. Findings from this paper shows that transformers gives more precision and recall values for the upto 400 words , if comparatively the number of words are more then word level CNN performs more accurately .The research gap is low signal to noise ratio , which means the the noise is less and data is less corrupted which makes easy for classification purposes.

**Paper 3**

**2023-**

**Automated Extraction of Pain Symptom in Electronic Health Records**

This paper tells about the clinical language annotation , modeling and processing , which enables recognition and automatic encoding of information. It uses word2vec ,and CLAMP model. The symptoms are categorized based on the words extracted and automatic encoding of information is performed .The formula used is Precision, recall, and F1 measure. The findings are as folows Pain quantity rating  is about  8/10 , and a good f1-score .The research gap is that the data is from single centre and also a pilot study , so there is less randomness in the data .

**3.1 Algorithms Used**

In the pursuit of accurate medical text classification, the selection of appropriate algorithms is pivotal. This chapter provides an in-depth exploration of the algorithms employed in the project, accompanied by detailed observations and insights into their performance.**Model 1: Using LSTM**Model 1 harnesses the power of Long Short-Term Memory (LSTM) networks, which are renowned for their ability to handle sequential data effectively. LSTMs are equipped with memory cells and gating mechanisms that enable them to capture long-range dependencies, making them well-suited for processing medical text inputs.**Observations**The performance of Model 1 was scrutinized across various configurations, including different optimizers, loss functions, and encoding types. Here are the key observations:**Optimizer Role:** The choice of optimizer plays a crucial role in optimizing the model parameters during training. Adam and RMSProp optimizers were tested, with Adam demonstrating superior convergence speed and stability.**Loss Function Significance**: The loss function determines the discrepancy between the predicted and actual outputs, guiding the optimization process. Categorical crossentropy emerged as the preferred choice, ensuring effective optimization by penalizing incorrect predictions.**Encoding Type Impact:** Encoding the output labels is essential for representing categorical data in a format suitable for neural network training. One-hot and two-hot encodings were experimented with, with one-hot encoding proving to be more effective in preserving categorical information and facilitating model convergence.**Performance Metrics**Model 1's performance was evaluated based on two key metrics: Test Loss and Accuracy. The results of the evaluations are as follows:With Adam optimizer and one-hot encoding, Model 1 achieved a remarkably low test loss of 0.0144, coupled with an impressive accuracy of 99.47%.Despite slight degradation in performance with different configurations, Model 1 maintained high accuracy levels, highlighting its robustness in medical text classification tasks.**Model 2: Using GloVe (ANN)**Model 2 integrates Global Vectors for Word Representation (GloVe) embeddings within an Artificial Neural Network (ANN) architecture. GloVe embeddings capture semantic relationships between words, enriching the feature space for classification tasks.**Observations**The performance observations for Model 2 are as follows:**Optimizer Impact:** The choice of optimizer significantly influences the model's convergence and generalization capabilities. Adam, Adagrad, and RMSProp optimizers were evaluated, with Adam showcasing the most promising results in terms of both convergence speed and final performance.**Loss Function Selection:** Sparse categorical crossentropy was chosen as the loss function for Model 2, given its suitability for multi-class classification tasks. This loss function effectively penalizes misclassifications while optimizing the model parameters.**Encoding Type Consideration:** Utilizing Label Encoder for encoding the output labels facilitated seamless integration with the ANN model, allowing for efficient representation of categorical data in a numerical format.**Performance Metrics**The performance evaluation of Model 2 yielded the following results:With Adam optimizer, sparse categorical crossentropy loss function, and Label Encoder encoding, Model 2 achieved a commendable test loss of 0.0244 and an impressive accuracy of 99.70%.Despite slight variations in performance across different configurations, all setups maintained exceptionally high accuracy levels, underscoring the efficacy of the GloVe-based ANN approach for medical text classification.In conclusion, both LSTM-based Model 1 and GloVe-based ANN Model 2 exhibit strong performance in medical text classification, showcasing the significance of algorithm selection and configuration in achieving accurate and reliable results.