SYMPTOM-BASED DISEASE PREDICTION CHATBOT USING DEEP LEARNING

TEAM MEMBERS:

- Lakshmi Mounika Bolisetti, Oam Chandra Laasya Tummala, Rakesh Reddy Yennam

1. INTRODUCTION

The landscape of healthcare diagnostics is undergoing a transformative revolution powered by machine learning technologies. In an era where early detection can significantly impact patient outcomes, our research addresses a critical gap in preliminary medical diagnosis. Traditional diagnostic methods often rely heavily on individual physician interpretation, which can be subjective and time-consuming. This project aims to develop an intelligent diagnostic support system that leverages the power of neural network architectures to provide rapid, accurate, and data-driven preliminary disease predictions.

2. PROBLEM DESCRIPTON

The complexity of medical diagnostics presents a multifaceted challenge that our research directly addresses. Symptom interpretation is inherently nuanced, with patients describing their experiences using varied language and medical professionals requiring comprehensive context to make accurate assessments. The primary goal was to develop a sophisticated predictive system that could transform raw symptom data into meaningful diagnostic insights. We aimed to create a solution that not only classifies potential diseases but also provides context-aware risk assessments and preventive recommendations. This approach goes beyond simple symptom matching, instead leveraging advanced machine learning techniques to identify complex patterns and relationships that might escape traditional diagnostic methods.

3. DATA DESCRIPTION

The dataset is derived from Kaggle [1], containing approximately 4,900 rows, which includes 132 symptoms and 41 diseases [2], is meticulously curated to include symptom descriptions, severity levels, precautionary measures, and disease mappings. Key components of the dataset are:

- **Symptom_severity.csv:** Details the severity levels of various symptoms on a standardized scale.
- **Symptom_precaution.csv:** Maps each disease to its recommended precautionary measures.
- **Training.csv:** Contains labeled data for training machine learning models, including disease-symptom relationships.
- **Symptom_Description.csv:** Offers comprehensive descriptions of symptoms to aid in natural language processing tasks.

4. METHODOLOGY

The methodology for our symptom-based disease prediction project employed a comprehensive machine learning approach, integrating advanced data preprocessing techniques with sophisticated neural network architectures.

4.1 Data Preprocessing and Preparation

Data preprocessing involved transforming raw symptom information into a machine learning-ready format. Symptoms were aggregated across columns, creating unified text representations and implementing robust missing value handling techniques. Disease labels were categorically encoded using scikit-learn's LabelEncoder, converting them into a numerical format suitable for neural network architectures. We employed a stratified train-test split, allocating 80%

of data for training and 20% for testing, with a fixed random state to ensure experimental reproducibility. This methodical approach laid the groundwork for developing an advanced diagnostic machine learning model.

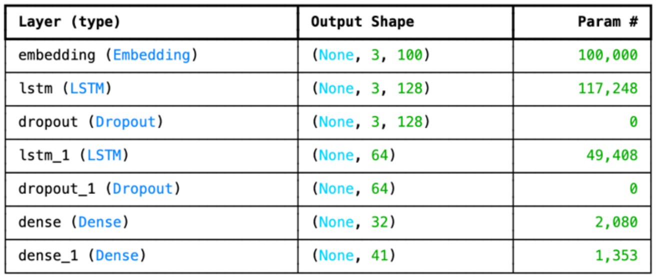
4.2 Text Tokenization and Sequence Preparation

Utilizing Keras Tokenizer, we converted raw symptom text into numerical sequences with a carefully selected maximum vocabulary of 1,000 words. An out-of-vocabulary token was incorporated to handle unseen words during prediction, ensuring the model's robustness when encountering novel symptom descriptions. Sequence normalization was achieved through padding techniques, standardizing input dimensions to a fixed length of three. This approach balanced the need for consistent input dimensions with the preservation of critical symptom information. Shorter sequences were padded with zeros, while longer sequences were truncated, creating a uniform input format for our neural network models.

4.3 Neural Network Architectures

Three distinct neural network architectures were developed and compared:

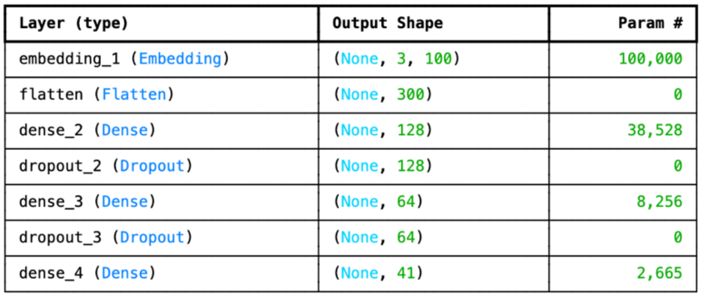
- *Long Short-Term Memory (LSTM) Network:*
  The LSTM model represented a sophisticated sequence processing architecture designed to capture temporal dependencies in symptom data. The model featured a multi-layered structure with an initial embedding layer transforming input sequences into dense vector representations. Two LSTM layers with 128 and 64 units were implemented, complemented by dropout layers to prevent neural overspecialization. L2 regularization was applied across layers to manage model complexity and prevent overfitting. Figure 1 demonstrates the model summary of the LSTM model used.



Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 3, 100)	100,000
lstm (LSTM)	(None, 3, 128)	117,248
dropout (Dropout)	(None, 3, 128)	0
lstm_1 (LSTM)	(None, 64)	49,408
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 32)	2,080
dense_1 (Dense)	(None, 41)	1,353

Figure 1: Model summary of LSTM model

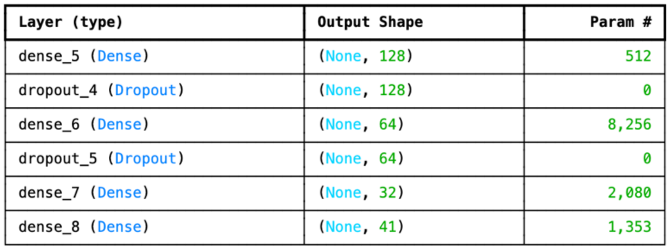
- *Artificial Neural Network (ANN):*
  The ANN approach adopted a dense layer architecture focused on complex feature extraction. Beginning with an embedding layer that was subsequently flattened, the network incorporated multiple fully connected layers with ReLU activation. Strategically placed dropout layers and L2 regularization ensured the model's generalizability while preventing overfit tendencies. Figure 2 shows the model summary of the ANN model used.



Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 3, 100)	100,000
flatten (Flatten)	(None, 300)	0
dense_2 (Dense)	(None, 128)	38,528
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 64)	8,256
dropout_3 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 41)	2,665

Figure 2: Model summary of ANN model

- *Feed-Forward Neural Network (FFN):*
  The FFN model represented a more traditional neural network approach, composed entirely of dense layers with ReLU activation. Dropout layers were integrated to mitigate overfitting risks, with a final softmax activation layer enabling multi-class classification. Model summary of feed forward network is shown is figure 3.



Layer (type)	Output Shape	Param #
dense_5 (Dense)	(None, 128)	512
dropout_4 (Dropout)	(None, 128)	0
dense_6 (Dense)	(None, 64)	8,256
dropout_5 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 32)	2,080
dense_8 (Dense)	(None, 41)	1,353

Figure 3: Model summary of Feed forward network

4.4 Model Training and Optimization

Model compilation utilized the Adam optimizer, selected for its adaptive learning rate and momentum capabilities. Sparse categorical cross-entropy served as the loss function, optimally suited for multi-class classification tasks. Each model underwent a comprehensive training process spanning 10 epochs, with 20% of the training data reserved for validation.

4.5 Performance Evaluation

The evaluation methodology extended beyond simple accuracy metrics. We implemented a comprehensive performance assessment strategy incorporating multiple evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC score. The classification reports provided granular insights into model performance across different disease classes, revealing nuanced strengths and limitations of each architectural approach.

4.6 Chatbot development

The model which will be used for the chatbot is selected based on performance evaluation, which will be discussed in **Section 5**. The system initializes a neural network model and loads symptom-disease datasets. Users input symptoms via a chatbot interface, where fuzzy matching handles typos or incomplete terms. Matched symptoms are tokenized, padded, and passed to the trained model, which predicts probable diseases with confidence scores. Disease details and precautions are retrieved from supplementary datasets and displayed. The chatbot maintains a conversation history for input refinement, ensuring an intuitive experience.

5. RESULTS

The ANN, as we can see in table 1, emerged as the most promising model, achieving an exceptional accuracy of 86.89%, significantly outperforming the LSTM and FFN models. The precision of the ANN reached 92.37%, indicating a highly reliable classification mechanism with minimal false positive predictions.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LSTM	0.8577	0.9006	0.8577	0.8323	0.9958
ANN	0.8689	0.9237	0.8689	0.8442	0.9960
FFN	0.6209	0.5218	0.6209	0.5482	0.9581

Table 1: Metrics comparison for all three models

The LSTM model substantiated its effectiveness with an accuracy of 85.77% and a precision of 90.06%. This model's architecture, specifically designed to capture sequential dependencies in symptom data, proved particularly adept at handling complex medical diagnostic scenarios. It's remarkable ROC-AUC score of 0.9958 underscored the model's exceptional discriminatory capabilities across various disease classifications.

In contrast, the Feed-Forward Neural Network demonstrated more modest performance, with an accuracy of 62.09% and precision of 52.18%. Despite these lower metrics, the FFN model maintained a surprisingly high ROC-AUC score of 0.9581, suggesting potential utility in specific diagnostic contexts.

Visualization techniques, including bar charts as shown in figure 4, further illuminated the models' performance characteristics. These graphical representations highlighted the nuanced differences in accuracy, precision, recall, and F1-score across the three neural network architectures. The consistent ROC-AUC scores above 0.95 for all models validated the robust nature of our machine learning approach.

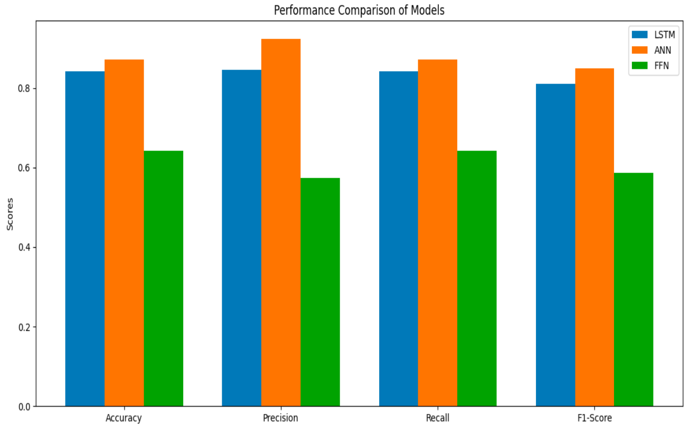


Figure 4: Bar plot visualizing the metrics of all three models

The confusion matrices provided granular insights into each model's classification behavior. Detailed analysis as shown in figure 5, revealed that the ANN and LSTM models exhibited minimal misclassification across disease categories, particularly excelling in distinguishing between symptomatically complex conditions. The Feed-Forward Network, while less precise, still demonstrated meaningful predictive capabilities.

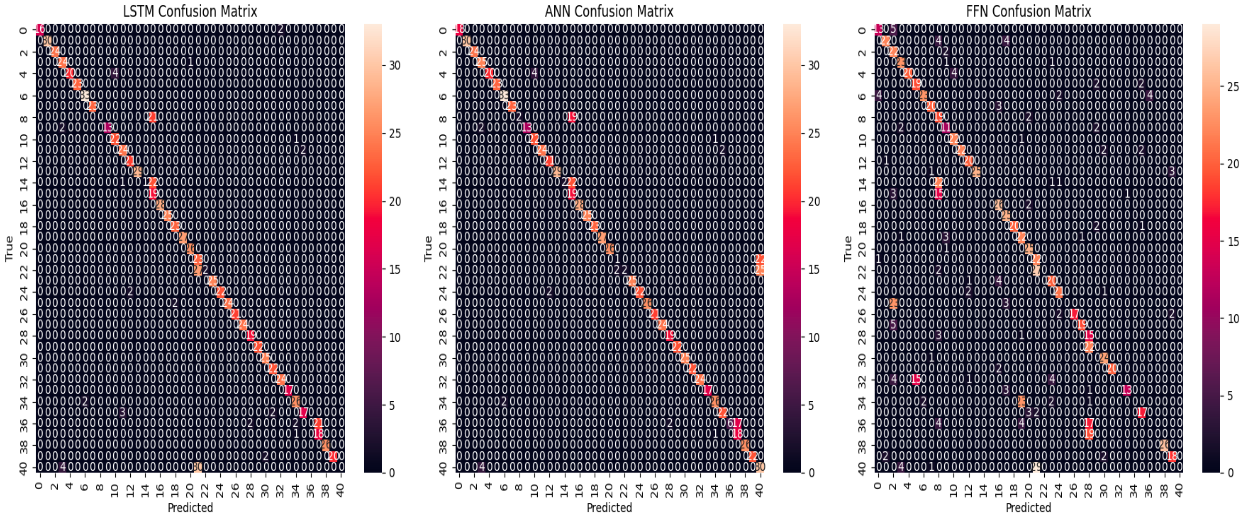
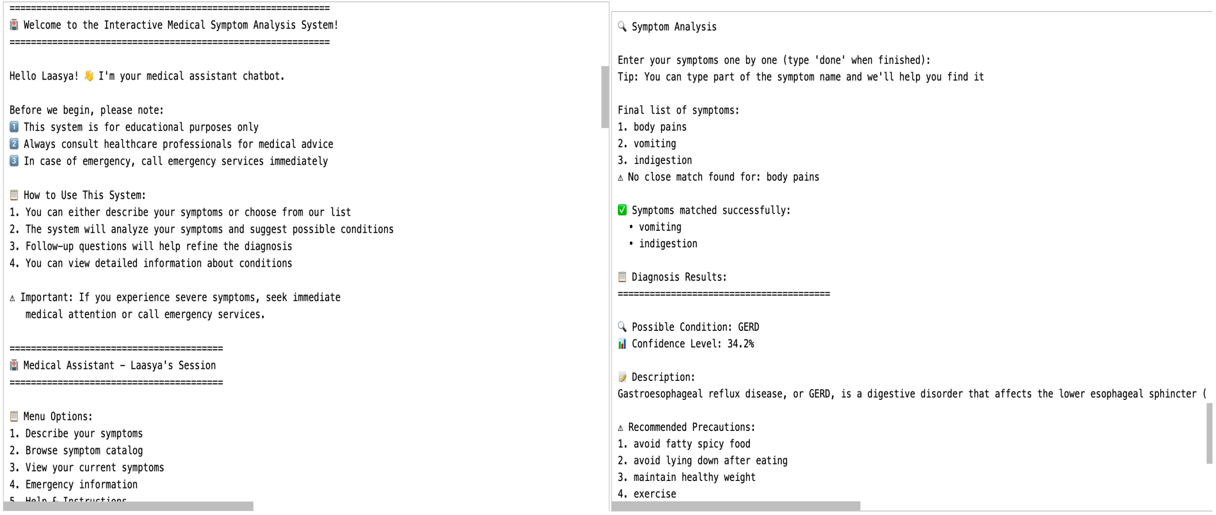


Figure 5: Confusion matrices for all three models

Ultimately, the ANN was selected as the primary model for the symptom-based disease prediction system chatbot. The interface and example response are shown in figure 6.



```
============================================================
🏥 Welcome to the Interactive Medical Symptom Analysis System!
============================================================

Hello Laasya! 👋 I'm your medical assistant chatbot.

Before we begin, please note:
1️⃣ This system is for educational purposes only
2️⃣ Always consult healthcare professionals for medical advice
3️⃣ In case of emergency, call emergency services immediately

📋 How to Use This System:
1. You can either describe your symptoms or choose from our list
2. The system will analyze your symptoms and suggest possible conditions
3. Follow-up questions will help refine the diagnosis
4. You can view detailed information about conditions

⚠ Important: If you experience severe symptoms, seek immediate
   medical attention or call emergency services.

========================================
🏥 Medical Assistant - Laasya's Session
========================================

📋 Menu Options:
1. Describe your symptoms
2. Browse symptom catalog
3. View your current symptoms
4. Emergency information
5. Help & Instructions
```

(a)

```
🔍 Symptom Analysis

Enter your symptoms one by one (type 'done' when finished):
Tip: You can type part of the symptom name and we'll help you find it

Final list of symptoms:
1. body pains
2. vomiting
3. indigestion
⚠ No close match found for: body pains

✅ Symptoms matched successfully:
  • vomiting
  • indigestion

📋 Diagnosis Results:
========================================

🔍 Possible Condition: GERD
📊 Confidence Level: 34.2%

📝 Description:
Gastroesophageal reflux disease, or GERD, is a digestive disorder that affects the lower esophageal sphincter (

⚠ Recommended Precautions:
1. avoid fatty spicy food
2. avoid lying down after eating
3. maintain healthy weight
4. exercise
```

(b)

Figure 6. (a) Interface of the chatbot, (b) Example response of the chatbot

6. DISCUSSION

The ANN and LSTM models demonstrated exceptional predictive performance, with ROC-AUC scores consistently exceeding 0.99, showcasing their ability to distinguish between different medical conditions. However, the ANN model outperformed the LSTM in terms of accuracy and computational efficiency, making it the preferred choice for deployment in this project. This performance difference can be attributed to the ANN's dense layer architecture, which effectively captured patterns in the structured, non-sequential dataset. In contrast, the LSTM model, designed to handle sequential dependencies, was less suited for this task and prone to overfitting due to the relatively small dataset size and limited temporal complexity.

The ANN model was integrated into a chatbot system designed to provide preliminary diagnoses and precautionary advice based on user-reported symptoms. This interactive system allows users to input symptoms in natural language, making diagnostic support accessible and user-friendly. While the model achieved high accuracy, there are still cases where predictions are incorrect. These errors often stem from ambiguities in how symptoms are described, insufficient representation of rare conditions in the dataset, or variability in symptom progression that the model cannot fully account for.

During development, several anticipated and experienced failures were encountered and addressed. One challenge was handling ambiguous and incomplete symptom inputs, which led to incorrect predictions. This was mitigated by implementing fuzzy matching techniques to match user-provided symptoms with those in the dataset. Another issue was the small dataset, consisting of only 132 symptoms and 41 diseases, which resulted in model overfitting and limited generalization. To address this, techniques such as dropout regularization and hyperparameter tuning were employed to improve model robustness. Additionally, computational limitations with the LSTM model were overcome by shifting focus to the ANN, which was more efficient for the given data.

7. FUTURE SCOPE

To address these limitations, future work will focus on expanding the dataset to include a broader range of symptoms and diseases, ensuring better representation of rare conditions. Real-time symptom tracking can be incorporated to analyze symptom progression and improve prediction accuracy. Furthermore, advanced models like BERT or GPT can be explored to better handle nuanced natural language inputs and improve the chatbot's interpretability. These enhancements will contribute to building a more robust and scalable system, capable of delivering personalized and accurate diagnostic insights.

8. REFERENCES

[1] https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

[2] Manikanta Sirigineedi, Matta Eswar Surya Manikanta Kumar, Rali Surya Prakash, Velagala Pavan Kumar Reddy, & Poojitha Tirunagari. (2024). Symptom-Based Disease Prediction: A Machine Learning Approach. *Journal of Artificial Intelligence,Machine Learning and Neural Network*, *4*(03), 8–17. https://doi.org/10.55529/jaimlnn.43.8.17

9. APPENDIX 1: TEAM CONTRIBUTIONS

- Lakshmi Mounika B: FFN, LSTM and ANN model development, Performance evaluation and Result analysis
- Oam Chandra Laasya Tummala: Data preprocessing and Chatbot development
- Rakesh Reddy: Data collection and management, Precautionary advice generation, documentation and reporting