**Diagnosing diseases using deep learning techniques**

Dr. Dina El-Sayad  
Scientific Computing DepartmentFaculty of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[dina.elsayad@cis.asu.edu.eg](mailto:dina.elsayad@cis.asu.edu.eg)

Malak Ismail   
Sicentific Computing Department  
Faculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700850@cis.asu.edu.eg](mailto:20201700850@cis.asu.edu.eg)

Mennattallh Ibrahim  
Sicentific Computing DepartmentFaculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700856@cis.asu.edu.eg](mailto:20201700856@cis.asu.edu.eg)

Manar Ibrahim  
Sicentific Computing Department  
Faculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700852@cis.asu.edu.eg](mailto:20201700852@cis.asu.edu.eg)

Manar Alaa   
Sicentific Computing DepartmentFaculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700853@cis.asu.edu.eg](mailto:20201700853@cis.asu.edu.eg)

Fady Makram  
Sicentific Computing Department  
Faculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700562@cis.asu.edu.eg](mailto:20201700562@cis.asu.edu.eg)

Manar Mohamed  
Sicentific Computing DepartmentFaculty Of Computers and information Sciences *Ain Shams University*Cairo, Egypt  
[20201700854@cis.asu.edu.eg](mailto:20201700854@cis.asu.edu.eg)

Abstract: The COVID-19 pandemic has driven a rise in remote healthcare solutions, including medical chatbots. These chatbots offer 24/7 access, faster service, and cost savings. However, Arabic chatbots pose unique challenges due to the language's complexity and dialect variations. This work introduces MAQA, the largest Arabic Healthcare Q&A dataset with over 430,000 questions across 20 specialties. Preprocessing techniques were applied for data refinement, followed by an 80/20 training-testing split. Four deep learning models Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Transformer, and AraBART-summ were evaluated for chatbot development. The AraBART-summ model achieved the highest average cosine similarity 96.6% and a BLeU score of 63%, demonstrating its potential for effective Arabic medical chatbots.

Keywords: Medical Chatbots, Natural language processing, Deep Learning, Transformer Model, MAQA Dataset

# **Introduction**

Limited access to healthcare facilities creates vulnerabilities in communities. This lack of proximity to essential medical services can lead to delayed treatment for illnesses or injuries, potentially resulting in worsened health outcomes and preventable mortality [[1](#ref1)]. Furthermore, these "healthcare deserts" exacerbate disparities in access, disproportionately impacting marginalized populations who already face barriers to care [[2]](#ref2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Author | Year | Approach | Accuracy | Similarity | BLeU |
| Mohammed Abdelhay et al [[1](#ref1)] | 2023 | LSTM  Bi -LSTM  Transformer | - | 56  72  80 | 31  39  58 |
| Abdennour Boulesnane et al [[2](#ref2)] | 2022 | LSTM, GRU  BI-LSTM | 78  71 | - | - |
| Qiming Bao et al [[3](#ref3)] | 2020 | BERT  MaLSTM  HBAM | 78.2  78.4  81.2 | - | - |
| Prateek Mishra et al  [[4](#ref4)] | 2022 | Rasa Architecture | 78.7 | - | - |
| Mohamed Boussakssouet al [[5](#ref5)] | 2022 | LSTM  GRU | 89  85 | Table 2 : Literature studies on deep learning papers  - | - |

Misdiagnosis presents significant threats, particularly when stemming from limited physician experience or information. Overlooked or misidentified conditions can delay appropriate interventions or lead to unnecessary procedures [[3](#ref3)]. Lack of expertise or comprehensive resources can hinder accurate diagnosis of complex or rare cases, negatively impacting individual patients and eroding trust in the healthcare system [[4](#ref4)].

The COVID-19 pandemic dramatically disrupted human interaction, with a significant decrease in in-person healthcare encounters. Hospitals became overwhelmed with COVID-19 patients, leading some individuals to avoid seeking care for unrelated conditions due to fear of contracting the virus [[5](#ref5)]. This reluctance to seek timely medical attention tragically contributed to avoidable deaths.

# **Related Works**

Traditional diagnostic methods, while instrumental, are susceptible to human error and limitations in interpreting complex medical data. Machine learning (ML) and deep learning (DL) are emerging as powerful tools in healthcare, potentially overcoming these limitations. These techniques can analyze vast datasets to identify subtle patterns and relationships that may elude traditional methods.

**Prior Research on Machine Learning and Deep Learning for Diagnosis**

Several studies have explored the application of ML and DL for disease diagnosis. These studies, summarized below, investigate diverse diseases and assess the effectiveness of different algorithms. Notably, some studies have shown promise in utilizing ML algorithms for early cancer detection and DL techniques for analyzing medical images. Tables [1](#Table1) and [2](#Table2) provide a more detailed overview of this prior research.

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Year | Approach | Accuracy |
| Asad Ur Rahman et al [[6](#ref6)] | 2022 | SVM | 93.33 |
| Sgar Badlani et al [[7](#ref7)] | 2021 | Random Forest | 98.43 |

Table 1 : Literature studies on traditional papers

# **Materials**

**A. MAQA Arabic Healthcare Q&A Dataset**

**Source:** The MAQA dataset was collected from various Arabic healthcare websites (e.g., altibbi.com, tbeeb.net, cura.healthcare). It is the largest publicly available Arabic Healthcare Q&A dataset to date.

**Description:**

* Approximately 430,000 questions across 20 medical specialties (detailed in Table [3](#Table3)).
* Unique, raw questions cleaned for basic formatting but not stemmed or lemmatized. Questions and answers may contain English symbols and digits but lack most Arabic diacritics and punctuation.
* Question content (q\_body), answer content (a\_body), word counts for both, category name, and category ID.
* Table [4](#Table4) displays a sample question from the "Gynecological" (امراض نسائية) and "Gastrointestinal" (امراض الجهاز الهضمي) categories.

|  |  |
| --- | --- |
| Label | Count |
| امراض نسائية | 103683 |
| امراض المسالك البولية والتناسلية | 33847 |
| امراض العضلات والعظام و المفاصل | 33050 |
| الامراض الجلدية | 29262 |
| الطب العام | 26870 |
| امراض باطنية | 23722 |
| امراض الجهاز الهضمي | 22373 |
| الامراض الجنسية | 21773 |
| طب الاسنان | 20207 |
| امراض الاطفال | 18636 |
| امراض نفسية وعصبية | 18295 |
| امراض القلب و الشرايين | 15368 |
| جراحة عامة | 15185 |
| امراض العيون | 14439 |
| انف اذن وحنجرة | 13933 |
| الاورام الخبيثة والحميدة | 11210 |
| امراض الغدد الصماء | 5186 |
| امراض الجهاز التنفسي | 4567 |
| جراحة تجميل | 1596 |
| امراض الدم | 1341 |

*Table 3: Categories and number of its record*

|  |  |  |
| --- | --- | --- |
| Question | Answer | category |
| اذا كانت مده الدورة عندي 32 يوم فماهي ايام الإباضة عندي لان بدي احمل | 14 15 16 17 18 | امراض نسائية |
| السلام عليكم  جاي يصير عده تهيج في تحدت عده خربطه في المعدة والقالون العصبي تستخدم علاج مرفق في الفايل ومستقرة على العلاج فقط تحدث مشكله في الاكل ماذا تكل وماذا... | This is not colon treatment it is stomach treatment | امراض الجهاز الهضمي |

*Table 4: examples of data*

**B. Disease Symptom Prediction Dataset**

**Description**: This English dataset serves as a resource for developing disease prediction or healthcare systems. It encompasses information on diseases, associated symptoms, recommendations, and weights. The data is structured around 41 distinct diseases and 17 symptoms. Table [5](#Table5) provides a sample entry from this dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Symptom | Disease | Description | precaution |
| الم المفاصل  وجع البطن  فقدان الشهية  القىْ | التهاب الكبد أ | التهاب الكبد (أ) هو عدوى الكبد شديدة العدوى الناجمة عن فيروس التهاب الكبد (أ). يعد الفيروس واحدًا من عدة أنواع من فيروسات التهاب الكبد التي تسبب الالتهاب وتؤثر على قدرة الكبد على أداء وظائفه. | استشارة أقرب مستشفى  غسل اليدين  تجنب الأطعمة الغنية بالتوابل الدهنية |

Table 5: Example of second data

# **Method**

## **A. System Architecture**

**1. Data Preprocessing**

Arabic text processing presents inherent challenges due to ambiguity, diglossia (variation between spoken and written forms), and complexities in script interpretation. Normalization techniques address these issues by rectifying inconsistencies in letter usage, dialect variations, and diacritics (symbols added to Arabic letters).

To overcome these challenges, Natural Language Processing (NLP) tools are employed. These tools include tokenization (splitting text into words), stemming (reducing words to their base form), and morphological analysis.

The data undergoes rigorous cleaning to eliminate duplicates, missing entries, and ambiguous sentences. Mislabeled data is corrected, and normalization techniques are applied for standardization.

Exploratory Data Analysis (EDA) helps understand the data by identifying symptom-related phrases within each category. This analysis also involves visualizing frequently used terms and calculating sentence lengths to determine the maximum input sequence length for models. This approach aids in detecting trends, patterns, and relationships within the Arabic data.

* **2. Data Augmentation**
* Data augmentation is a technique that injects additional data or priors to improve model performance and prevent overfitting. The paper discusses various augmentation methods:
* **Symbolic Augmentation:** This method utilizes auxiliary neural networks or statistical models for data generation. While interpretable, it has limitations in complex transformations.
* **Rule-based Augmentation:** Pre-defined rules are employed to create augmented examples (e.g., synonym replacement, random deletion).
* **Back-translation Augmentation:** Text is translated to another language and back to introduce variations.
* **Style Augmentation:** Deep networks transfer writing styles between authors (e.g., for abstractive summarization).
* **Generative Data Augmentation:** This method involves generating realistic text passages for training deep learning models.
* **3. Text Preprocessing Techniques**
* **Tokenization:** This process breaks down user queries into a list of words for easier processing.
* **Lemmatization:** Tokenized data is normalized through lemmatization to enhance model learning (e.g., converting "running" and "runs" to "run").
* **4. Feature Extraction**
* Following tokenization and lemmatization, the BoW model converts text data into numerical vectors suitable for model training. The model identifies words related to specific medical conditions based on the processed text.
* Pretrained Language Modelslike BERT, GPT, and RoBERTa are employed for text preprocessing to enhance accuracy and efficiency. These models perform tasks like tokenization and context-aware word embedding, reducing computational demands.
* **5. Classification System**
* The system employs machine learning techniques for medical specialty identification, a crucial step in early diagnosis. This classification automates administrative processes and aids medical specialists in managing patient treatment pathways.
* **6. Assistant System**
* This study explores the use of AI-powered chatbots to improve healthcare practices and develop cost-effective resources. Advancements in NLP enable chatbots to automate conversations and respond promptly to user inquiries.

## **B. System implementation**

The chatbot development process can be divided into three key stages: preprocessing, classification, and text generation. Each stage employs specific techniques to optimize chatbot quality and functionality.

1. **Preprocessing**

**a)** **Data Balancing:**

Imbalanced datasets, where certain medical specialties are overrepresented, can be addressed using techniques like translation-based augmentation, word swapping, and data scraping.

**b)** **Classification Preprocessing:**

1. This stage prepares the data for classification tasks:
2. Irrelevant data columns are removed (e.g., a\_body, a\_bodycount).
3. The maximum question length is determined.
4. Labels are encoded for categorical representation.
5. Questions are tokenized (split into individual words).
6. Question bodies are padded to a uniform length for model processing.

**c)** **Generation Preprocessing:**

1. This stage prepares the data for response generation:
2. The average question length is used as the target sequence length for generating responses.
3. Duplicate entries and missing values are identified and addressed.
4. Extraneous content like advertisements is removed.
5. Diacritics (optional) may be removed to simplify text representation, considering potential accuracy impact.
6. Repetitive characters are eliminated for normalization.
7. Arabic letters are standardized for consistency.
8. Numerical digits are converted to words for improved language comprehension.
9. English text is identified and handled (e.g., machine translation or removal) to maintain coherence.
10. Data filtration sets a maximum length limit for questions and answers, retaining relevant information.
11. Answer trimming restricts response word count for conciseness.
12. Word counts for both questions and answers are generated.
13. A vocabulary is defined based on word frequency thresholds, including special tokens.
14. Text is encoded by converting cleaned questions to their corresponding vocabulary indices. Sequences are padded or truncated to meet the maximum length requirement.

**Models**

***First classification models***

## **Machine learning models**

**Random Forest:**

An ensemble learning method that combines predictions from multiple decision trees, improving overall accuracy and robustness. Random forests are efficient for handling high-dimensional data and can provide interpretability through feature importance scores.

**Multinomial Naive Bayes (MNB):**

A probabilistic classifier based on Bayes' theorem, assuming independence between features. It's particularly well-suited for text classification with discrete features like word counts or term frequencies. MNB offers efficient training and classification but may not capture complex relationships within the data.

**Support Vector Machine (SVM):**

A powerful supervised learning algorithm for both classification and regression tasks. SVM finds the optimal hyperplane that separates data points into distinct classes, maximizing the margin between them. SVMs are well-suited for high-dimensional data and can handle small datasets effectively. However, they may require careful hyperparameter tuning and can be less interpretable compared to other models.

## **Deep learning models**

Deep learning models offer advanced capabilities for text classification by capturing complex relationships within data. Here, we explore several prominent architectures:

* **Long Short-Term Memory (LSTM) Networks:** A type of recurrent neural network (RNN) capable of learning long-term dependencies. LSTMs are well-suited for sequence prediction problems like text classification. Their internal memory cells allow them to store and access relevant information across long sequences.
* **Transformer Model with Gated Recurrent Unit (GRU) for Text Classification:** This model combines the strengths of Transformers, adept at capturing global dependencies, and GRUs, effective for temporal dependencies within sequences. Additionally, early stopping helps prevent overfitting.
* **Fine-tuned BERT Model (Bidirectional Encoder Representations from Transformers):** This approach leverages a pre-trained BERT model, known for its state-of-the-art performance in natural language understanding tasks. Utilizing the AutoModelForSequenceClassification and AutoTokenizer functions from the Hugging Face Transformers library to fine-tune a pre-trained BERT model on the specific medical domain classification task. Fine-tuning allows the model to adapt its pre-trained knowledge to the new problem.

***Second generation models***

**Seq2Seq models with Long-Range dependencies:**

**LSTM (Long Short-Term Memory):** a specific type of Recurrent Neural Network (RNN), address the limitations of traditional RNNs in handling long-term dependencies within sequential data. LSTMs utilize memory cells to store information for extended periods and employ gates to control information flow within these cells. Trained via gradient descent and backpropagation, LSTMs excel at text generation by iteratively predicting the next word based on context, making them a powerful tool for various Natural Language Processing tasks.

**Bi-LSTM (A Bidirectional LSTM):** extends the standard LSTM by processing input sequences in both forward and backward directions.

It consists of two LSTM layers: one processes the sequence from left to right, and the other from right to left. This bidirectional processing allows the model to capture both past and future dependencies for each token. The forward LSTM layer processes the input in a standard sequential manner, while the backward LSTM layer processes it in reverse order. As both layers process the sequence, they maintain hidden states that encode the learned information. After processing, the hidden states from both directions are concatenated, providing a comprehensive representation of the input sequence.

The Bi-LSTM model is trained using backpropagation and gradient descent, adjusting weights and biases to minimize prediction errors. For text generation, the model takes an initial input and iteratively predicts the next word or character based on the combined context from both directions. The advantages of Bi-LSTM include capturing bidirectional dependencies and enhanced context understanding, leading to richer text generation. Bi-LSTM models are effective in various natural language processing tasks, such as text generation, sentimentanalysis, and named entity recognition, offering improvedcontextunderstanding compared to unidirectional LSTMs. However, their use should consider computational complexity and potential overfitting.

In recent advancements, the performance of LSTM and Bi-LSTM models has been significantly enhanced by integrating them with attention mechanisms this integration allowed us to approach better result

**Transformers**: introduced by Vaswani et al. [[8](#ref8)], have revolutionized NLP tasks like text generation. Their secret lies in a multi-layered structure that utilizes self-attention mechanisms alongside feed-forward networks. The self-attention layer analyzes the input sequence, focusing on relevant parts and capturing both local and long-range dependencies. This refined information is then processed by feed-forward layers for further enhancement. Transformers excel due to their ability to efficiently process information in parallel, handle large datasets, and offer interpretability through the attention mechanism. These advantages make them a powerful and adaptable architecture for various NLP applications.

**disease-symptom-prediction**

**Preprocessing**

1. Initially, English data was translated to Arabic using Google Translate.
2. Data was shuffled to remove order bias
3. Cleaning by replacing underscores and filling missing values with zeros.

**Models**

This study investigated the application of machine learning models for text classification tasks. Three prominent algorithms were employed: Random Forest, Decision Tree, and Support Vector Machine (SVM).

* **Random Forest:** This ensemble method is known for its ability to handle high-dimensional data, such as those encountered in text classification. Random Forests excel at capturing complex relationships between features, leading to robust performance.
* **Decision Tree:** Decision Trees offer a simple yet powerful approach to classification. They work by constructing a tree-like structure where each node represents a decision based on a specific feature. The data is recursively split at each node based on the feature that best separates the classes. This process continues until a stopping criterion is met, resulting in a set of terminal nodes (leaves) representing the predicted class labels.
* **Support Vector Machine (SVM):** SVMs are powerful classifiers adept at learning complex patterns and effectively discriminating between classes. This makes them well-suited for text classification tasks where data can exhibit intricate relationships.
* **Ensemble Voting:**

The final model employed a voting ensemble approach. This strategy combines the predictions of multiple models (in this case, Random Forest, Decision Tree, and SVM) to potentially improve overall classification accuracy. Each individual model votes on the predicted class label, and the final prediction is determined by the majority vote. This approach can leverage the strengths of each individual model and mitigate potential weaknesses.

# **Model Evaluation**

The deep learning model uses trained word vectors from the pre-trained CBOW model called Aravec, trained on 132,750,000 Arabic documents with 3,300,000,000 words. A word embedding matrix is generated from this model, and the word embedding sentences of the corpus are fed to the network as input features. The model is tested on a held-out test set and evaluated using metrics such as cosine similarity and BLEU score as following:

## Cosin Similarity

To evaluate our generated answer against the actual answer, we start by getting the embedding vector for each word in the sentence, then get the average for all words’ vectors as in equation A1, where A is the sentence vector, Vi is the word vector and N the words count in the sentence. Then, we calculate the vectors product in equation A2, where A and B are two nonzero vectors can be derived by using the Euclidean dot product formula. After that we calculate the cosine similarity between both average vectors as in equation A3, where Ai and Bi are its components of vectors A and B, respectively. Finally, we calculate the Cosine Distance as in equation A4. The greater Cosine Distance, the greater the model efficiency and accuracy (Hendy et al. 2023) [[9](#ref9)]

*Cosine Similarity Equations*

A diagram of mathematical equations

Description automatically generated

(A1)

(A2)

(A3)

(A4)

## BLEU Score

The BLeU score is an algorithm for evaluating the quality of text generated using deep learning algorithms (Papineni et al. 2002) [[10](#ref10)]; accuracy is considered the correspondence between a machine’s output and that of a human. The base stone of the BLeU score is the familiar precision measure, which is calculated by counting the number of candidate translation words (unigrams) that occur in any reference translation and then divided by the total number of words in the candidate translation as shown in equation A5 (Hendy et al. 2023) [[9](#ref9)]. However, as in our bot task, the modified n-gram can be generalized as in equation A6 to the case: one candidate sentence and one reference sentence, where ŷ is candidate sentence and y is one reference sentence. Then, we start with the n-gram count summation as in equation A7 (Hendy et al. 2023) [9]. This count summation cannot be used to compare sentences since it is not normalized. If both the reference and the candidate sentences are long, the count could be huge, even if the candidate is of poor quality (Hendy et al. 2023) [[9](#ref9)]. So, we normalize it as in equation A8, and equation A9 shows the final definition of BLEU, where w ∶= (w1, w2, ⋯) is the weighting vector, and Ŝ ∶= (ŷ (1), ⋯, ŷ(M)) is candidate corpus, and S = (S1, ⋯, SM) is reference candidate corpus

A close-up of a math problem

Description automatically generatedA black text on a white background

Description automatically generated*BLeU Score Equation*

(A1)

(A2)

(A3)

(A4)

(A5)

A black text on a white background

Description automatically generated

A close up of a math equation

Description automatically generatedA black text with a black line

Description automatically generated

# 

# **Results**

## **Results of Unbalanced Data**

On the MAQA dataset with imbalanced classes, we evaluated the performance of LSTM, Bi-LSTM, and Transformer models. The Transformer achieved the highest average similarity score (91.8%) and a promising average BLEU score (63.1%).

**figure 2. results of unbalanced data**

Analysis of the dataset revealed significant class-dependent performance variations. The "أمراض الجهاز التنفسي" (respiratory diseases) class achieved the highest performance (similarity score: 96.65%) with 8,764 records. This success is likely attributable to model parameters specifically optimized for this class. Conversely, the "أمراض نسائية" (gynecological diseases) class, despite having the most records (70,160), exhibited the lowest performance (similarity score: 90.2%). This suggests potential suboptimal parameterization for this class.

For the MAQA run on the entire dataset, we utilized the Transformer model due to its superior performance, achieving an average similarity of 83.8% and BLEU score of 59.58%.

## **Results of Balanced data**Top of Form

Our approach utilizes class-specific optimization for text generation. We iteratively process each class, adjusting model parameters based on two factors: total data volume within the class file and average text length. This optimization aims to enhance model performance for each unique class.

Following parameter optimization, we perform classification to generate individual models. These models become suitable for generating text relevant to the class they were trained on.

For question answering, the predicted category of a question determines the appropriate model to load. This loaded model, either a fine-tuned model or a Transformer, leverages its class-specific knowledge to generate a response relevant to the question's scope.

Our evaluation revealed a slight performance difference between the employed generation models. The fine-tuned model achieved a similarity score of 96.6%, while the Transformer model yielded a score of 95.4%. However, the fine-tuned model exhibited a higher BLEU score 63% compared to the Transformer's 61.5%.

**figure 3: Pre-classifier Result**

## **Results of all data**

For the run on the entire dataset, the Transformer model was chosen due to its superior performance in prior evaluations, as evidenced by its maximum average similarity and BLEU scores. This selection proved effective, achieving a similarity score of 96.2% and a BLEU score of 63%.

|  |  |  |
| --- | --- | --- |
| Data | Models | |
| **Unbalanced Data** | **Transformer**  Similarity :83%  Bleu: 59.9% | |
| **Balanced Data** | **Transformer**  Similarity :95.7%  Bleu: 62% | **Fine Tuning**  Similarity :96.2%  Bleu: 63% |

*Table 6: compare between balanced & unbalanced data*

## **Diagnosing Expermental results**

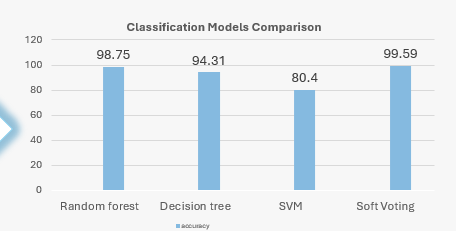
This study investigated machine learning for disease prediction. We split data (80% training, 20% testing) represented using embedding features. Various models were evaluated in Table [7](#Table7).

To potentially improve performance, a hard voting ensemble combining these models was implemented. This ensemble achieved a superior accuracy of 99.95%.

Following disease prediction, the system retrieves the corresponding disease description and symptom precautions from designated files. For optimal accuracy, users must enter at least four symptoms, allowing for more robust disease classification.

|  |  |
| --- | --- |
| Model | Evaluation |
| Random Forest | **accuracy 98.75%**  **F1 Score 98.73 %**  **Precision 98.8%**  **Recall 98.75%** |
| Decision Tree | **accuracy 94.31%**  **F1 Score 95.38%**  **Precision 96.7%**  **Recall 95.5%** |
| SVM | **accuracy 94.31%**  **F1 Score 95.38%**  **Precision 96.7%**  **Recall 95.5%** |
| Soft voting | **accuracy 94.31%**  **F1 Score 95.38%**  **Precision 96.7%**  **Recall 95.5%** |

*Table 7: compare between diagnosing ML models*

**figure 4: Diagnosis Results**

# **Discussion**

This section explores the challenges of Arabic's complexities and dialects for AI, particularly in chatbot development. Despite these hurdles, chatbots offer significant benefits, especially in healthcare communication.

Our project, "DIGNOSY," addresses this gap by creating a versatile Arabic chatbot (supporting dialects) for improved healthcare delivery. Built on sequence-to-sequence models (LSTM, Bi-LSTM, Transformer), DIGNOSY demonstrates promising results despite a limited training dataset.

Our analysis reveals a correlation between sentence length (both questions and answers) and chatbot effectiveness. This suggests potential improvements in handling longer user inputs.

**Overall, DIGNOSY shows promise in enhancing Arabic healthcare communication. Future work can focus on expanding the training data and exploring techniques for handling longer sentences. Additionally, DIGNOSY's architecture holds potential for applications beyond healthcare.**

# **Conclusion**

The intricate nature of Arabic, with its rich morphology and dialect variations, poses a significant challenge for AI, especially in chatbots. However, chatbots offer tremendous potential in healthcare, streamlining services and providing accessible support.

DIGNOSY, our proposed chatbot, aims to bridge this gap in Arabic healthcare communication. Utilizing powerful sequence-to-sequence models (LSTM, BiLSTM, Transformer), DIGNOSY converses with patients in their native dialect.

Evaluation demonstrates its efficacy in handling diverse queries despite a limited dataset of [insert actual data size] entries. Interestingly, analysis reveals a correlation between sentence length (both user input and responses) and chatbot effectiveness.

# **Future Work:**

This research establishes a foundation for further development. Future efforts will target:

* **Enlarged Training Data:** Expanding the dataset will enhance DIGNOSY's accuracy and ability to handle complex medical inquiries.
* **Medical Knowledge Integration:** Integrating domain-specific medical knowledge will enable more comprehensive, evidence-based responses.
* **User Feedback Mechanisms:** Implementing user feedback mechanisms will facilitate continuous improvement of DIGNOSY's performance and user experience.

By addressing these areas, DIGNOSY has the potential to become a valuable tool in Arabic healthcare. It can empower patients with readily accessible health information and improve communication between patients and healthcare providers.

# **Acknowledgment**

The successful completion of this project would not have been possible without the contributions of several individuals.

First and foremost, I express my deepest gratitude to Allah for granting me the strength and guidance to see this work through.

I am incredibly grateful to my parents and family for their unwavering support and encouragement throughout my academic journey. I strive to make them proud.

My sincere appreciation goes to my esteemed supervisors, Prof. Dr. Dina El-Sayad and T.A. Radwa El-Hussieny. Their patience, knowledge, and invaluable guidance were instrumental in shaping this project.

Finally, I extend my thanks to my friends and all those who offered their support and encouragement throughout this endeavor.

# **References**

1. Abdelhay, Mohammed, Ammar Mohammed, and Hesham A. Hefny. "Deep learning for Arabic healthcare: MedicalBot." *Social Network Analysis and Mining* 13.1 (2023): 71.
2. Boulesnane, Abdennour, et al. "Dzchatbot: a medical assistant chatbot in the algerian arabic dialect using seq2seq model." *2022 4th international conference on pattern analysis and intelligent systems (PAIS)*. IEEE, 2022.
3. Bao, Qiming, Lin Ni, and Jiamou Liu. "HHH: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention." *Proceedings of the Australasian computer science week multiconference*. 2020.
4. Mishra, Prateek, et al. "Personalized Healthcare Chatbot: Dataset and Prototype System." *International Conference on Computational Intelligence in Communications and Business Analytics*. Cham: Springer International Publishing, 2022.
5. Chatbot in Arabic language using seq to seq model

[https://www.researchgate.net/publication/355961405\_Chatbo in\_Arabic\_language\_using\_seq\_to\_seq\_model](https://www.researchgate.net/publication/355961405_Chatbo%20in_Arabic_language_using_seq_to_seq_model)

1. Health Consultant Bot: Primary Health Care Monitoring Chatbot for Disease Prediction <https://journal.50sea.com/index.php/IJIST/article/view/193>
2. Multilingual Healthcare Chatbot Using Machine learning

<https://www.researchgate.net/publication/352668726_Multilingual_Healthcare_Chatbot_Using_Machine_Learning>

[8] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.

[9] Hendy, Amr & Abdulrahim, Mohamed & Sharaf, Amr & Raunak, Vikas & Gabr, Mohamed & Matsushita, Hitokazu & Kim, Young Jin & Afify, Mohamed & Awadalla, Hany. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. 10.48550/arXiv.2302.09210

[10] Papineni K, Roukos S, Ward T, et al (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318, [https://doi.org/10.3115/10730 83.1073135](https://doi.org/10.3115/10730%2083.1073135)