An Ensemble Classification Approach in A Multi-Layered Large Language Model Framework for Disease Prediction

*Note: Sub-titles are not captured for https://ieeexplore.ieee.org and should not be used

1st Given Name Surname
Department Name
Affiliation/Organization
City, Country
email@example.com or ORCID

2nd Given Name Surname
Department Name
Affiliation/Organization
City, Country
email@example.com or ORCID

3rd Given Name Surname
Department Name
Affiliation/Organization
City, Country
email@example.com or ORCID

***Abstract*—Social telehealth has made a remarkable progress in the healthcare by allowing patients to post symptoms and participate in medical consultations remotely. Users frequently post symptoms on social media and online health platforms, creating a huge repository of medical data that can be leveraged for disease classification. Big models such as LLAMA3, GPT-3.5 Turbo, Furthermore, BERT processes intricate medical data, thereby enhancing disease. Classification. The current study compares three Arabic medical texts. preprocessing methods: text summarization, text refinement and Named Entity Recognition (NER). We utilized an ensemble majority voting method by combining the outputs of Named Entity Recognition (NER), as well as summarized, refined, and original posts, utilizing three Arabic pretrained models: CAMeLBERT, AraBERT, and AsafayaBERT. This ensembling approach achieved the best classification accuracy of 80.56%, thus showing its effectiveness in leveraging various text representations and model predictions to improve the understanding of medical texts.**

***Index Terms*—Text Classification; Social Tele-Health; Large Language Models; Natural Language Processing**

I. Introduction

The growth of social telehealth has revolutionized the provision of healthcare, enabling patients to share their symptoms and even consult with doctors remotely . During the COVID-19 pandemic, social telehealth increased greatly in popularity; at that time, access to traditional healthcare services was limited. Social media platforms, particularly health forums, are becoming increasingly valuable sources of user-generated medical data that people use to share detailed symptoms and seek the advice of doctors' opinions [1]–[3]. However, the unstructured and noisy nature of this data poses a great challenge; hence, advanced computational techniques are required for effective analysis. Recent advances in NLP, especially with the advent of large language models, have really transformed unstructured textual data processing [1]–[3]. Various models, including LLAMA, GPT, and BERT, among others, had promised outstanding performance related to text classification and understanding based on large-scale pre-training over big datasets to obtain advanced on a wide range of domains [3]–[6]. With these advances, challenges arise in applying LLMs to domain-specific tasks, including healthcare. The fine-tuning of specialized application LLMs have to consider domain-specific nuances, noisy data handling, and optimizing computational efficiency. This is a resource- and computationally intensive process involving new frameworks so that the effectiveness of LLMs in real-world applications is optimized [7]–[9].

We propose a framework that combines LLM-based preprocessing with fine-tuning of Arabic language models for disease classification. It refines text, summarizes posts, and extracts medical entities using NER, enhancing task-specific fine-tuning and overcoming traditional method limitations.

Figure 1 illustrates the flow process of our proposed framework. The raw text data generated by users is refined, summarised, NER, and annotated using LLM-based preprocessing. Further, the enriched dataset is utilized to fine-tune Arabic language models for multi-class, multi-label classification tasks, including disease type prediction.

Our contributions in this work are as follows:

- **Novel Integration of LLMs and Arabic Language Models:** We present a multi-layered framework that employs LLMs for preprocessing (refinement, summarization, NER) to enhance the fine-tuning of Arabic language models intended for healthcare applications.
- **Enhanced Preprocessing Pipeline:** We create an improved dataset. LLM-based preprocessing for making challenges related to user-generated content more interpretable and more structured.
- **Improvement of Classification Performance:** We enhanced the accuracies of disease type classification by combining fine-tuned pre-trained Arabic language models like CAMeL-BERT, AraBERT, and Asafaya-BERT on LLM-preprocessed data.

This work emphasizes the transformative potential of exploiting LLMs in order to enhance fine-tuning for domain-

Identify applicable funding agency here. If none, delete this.

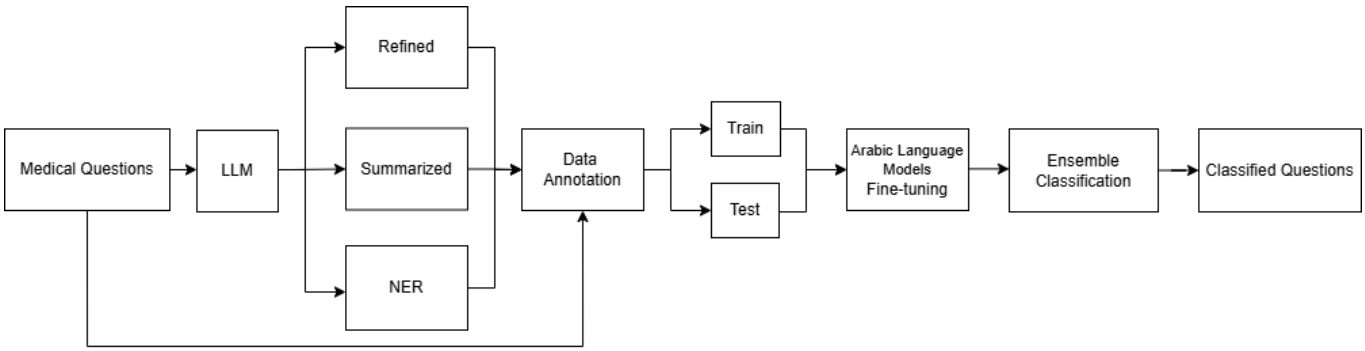

Fig. 1. Proposed Ensembling approach in a Multi-Layered Framework for Enhancing Arabic Language Model Fine-Tuning with LLAMA3 Preprocessing

specific language models when handling real-world challenges in healthcare.

II. Related Work

Recent breakthroughs in text classification and fine-tuning of large language Models have enhanced NLP applications across domains. Transformer-based models such as BERT and its variants have achieved remarkable performance on a variety of tasks such as false information detection, sentiment analysis and radical content classification [10]–[13]. Fine-tuning even on small labeled datasets improves performance and bidirectional context capture. Refinements like RoBERTa's NSP task removal boost domain-specific results, achieving an F1 score of 0.8 in medication detection and 3.929 MAE in eye-tracking tasks [12], [14], [15].

Efficient light-weight models like DistilBERT, which is 40% smaller and 60% faster, retain 97% BERT's performance, excelling in tasks such as the classification of socio-political news [12], [16]. ALBERT-Base-v2 optimized memory and speed, achieving a 13.04 exact match score in COVID-19 queries [12], [17]. XLM-RoBERTa improved results in processing more than 100 languages using the advances in multilingual processing and outperforming prior models by 23% in multilingual accuracy [12], [18]–[21]. Special adaptations like Electra-Small and BART-Large introduced token substitution and hybrid architectures. Electra-Small excelled in multilingual fake news detection with innovative token substitution strategies [12], [22], [23]. Both models achieved high performance in specialized tasks such as medical complaint detection and NER, with BART-Large reducing voice recognition errors by 21.7% [12], [24], [25]. These advancements highlight LLMs' potential in handling noisy and unstructured text data.

Generative LLMs enhance NLP, excelling in specialized tasks. Fine-tuned models like GPT-3.5 and Mistral-7B surpass baselines by 50% in F1-macro scores on domain-specific datasets [26]. QLoRA improves memory efficiency [27], but model variability persists, as seen in GPT-NeoX-20B and Llama2-7B.

Preprocessing techniques such as text refinement, summarization, and Named Entity Recognition (NER), are essential for handling noisy and unstructured text. BERT and RoBERTa have performed extremely well in NER tasks, effectively extracting key entities from complex text [9], [12], [14]. Combining preprocessing with fine-tuning enhances classification accuracy, especially when raw data lacks structure or clarity.

Based on this, our study introduces a multi-layered framework that uses LLM preprocessing to improve the fine-tuning of Arabic language models and address noise in data. This enhances the effectiveness of Arabic language models such as CAMeL-BERT, AraBERT, and Asafaya-BERT, in disease classification. It also presents an effective strategy for incorporating LLMs into social telehealth applications.

III. Dataset

The dataset used in this study was collected from user-generated posts on an online social platform where patients shared their medical complaints in Arabic. The posts contained detailed information such as the status of chronic diseases, descriptions of symptoms, symptom durations, height, weight, gender, and age, categorized into Type, and Diagnosis. Structuring and annotation were done under the supervision of a medical advisor to ensure relevance and accuracy.

Figure 2 shows the dataset's distribution, with Figure 2 illustrating the variety of medical issues like chronic diseases, skin conditions, and neurological symptoms.

IV. Methodology

This work proposes an Ensemble Classification on Arabic language model fine-tuning using a multi-layered framework. The LLAMA3 model has been used in the enhanced preprocessing step; the flow is illustrated as shown in Fig. 1.

A. Medical Questions

The dataset contains Arabic text data from user-generated content on online health platforms. Texts include elaborate descriptions by the patients of their symptoms, patient medical history, age, and gender. While doing preprocessing, it automatically removes private and sensitive information to ensure privacy.

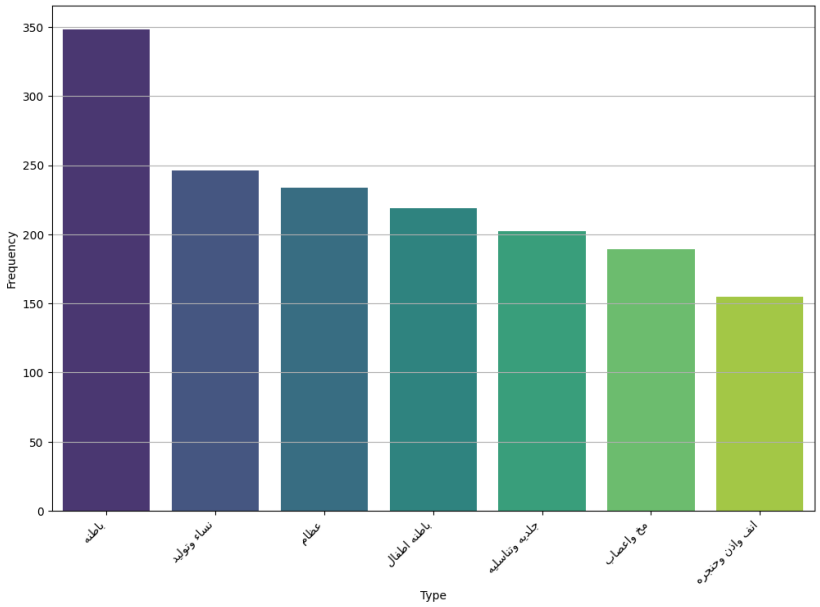

Fig. 2. Distribution of condition types in the dataset, illustrating the diversity of medical issues represented.

B. Multi-layer preprocessing using LLAMA3

The text is further enhanced using a multi-layered approach to LLAMA 3 filtering. The process involved in preprocessing is as follows:

- **Text Refinement:** LLAMA3 improves the text through deleting unmet requirements such as irrelevant information, grammatical mistakes and the vague parts of the material while protecting the medical context.
- **Text Summarization:** Llama 3 eliminates unnecessary aspects of long medical posts and compresses them into summaries for easier understanding.
- **Named Entity Recognition (NER):** Symptoms and conditions, together with the drugs, are important medical entities and LLAMA 3 identifies and extracts them.

All these steps (refined text, summarized text, and NER-extracted entities) are used to augment the base data.

C. Data Annotation

To create new data variants, the performance data is combined with the outcomes of the preparation stages. These datasets are effective in multi-label classification tasks:

- **Diagnosis Classification:** The goal is to identify the particular medical condition.
- **Type Classification:** This deals with the categorization aspect of the condition (e.g., chronic, acute).

D. Arabic Language Models Fine-Tuning

Three different pre-trained transformer models were fine-tuned individually for the classification task:

- **CAMeL-Lab/bert-base-arabic-camelbert-mix**
- **aubmindlab/bert-base-arabert**
- **asafaya/bert-base-arabic**

Each model was adapted by adding a dropout layer followed by a linear classification head to output predictions over the seven categories. All models were trained under the same hyperparameter configuration to ensure a fair comparison. The training configuration was as follows:

- Dropout Rate: 0.05
- Learning Rate: 1e-4
- Batch Size: 4
- Number of Epochs: 25
- Weight Decay: 0.01
- Loss Function: Cross-Entropy Loss

E. Ensemble Learning: Majority Voting

After fine-tuning the individual models, an ensemble strategy based on Majority Voting was applied. For each question in the test set, predictions from all three models were collected, and the final prediction was determined based on the majority class among the outputs.

The ensemble approach aimed to combine the strengths of different models, reduce variance, and improve the overall prediction robustness and reliability.

V. Results

To evaluate the classification performance of different ensemble configurations, we constructed ensemble combinations from twelve base models:

- **CAMeLBERT**: Post, Refined, NER, Summarized
- **AraBERT**: Post, Refined, NER, Summarized
- **AsafayaBERT**: Post, Refined, NER, Summarized

These models were selected due to their linguistic diversity and preprocessing variations, allowing us to test the effects of ensemble size and model variety.

TABLE I
COMPARISON OF TEXT, REFINED, SUMMARIZED, AND NER

Text	Refined	Summarized	NER
السلام عليكم ورحمة الله وبركاته شاب ٢٢ سنه بروح الجيم يوميا بس بحس أحيانا بضعف والم ف العضلات بعد التمرين محتاج دكتور مختص يردء عليا بعلاج مناسب لتقوية الأعصاب او فيتامينات لتقوية الأعصاب الوزن ٥٥ الطول ١٦٩	الشاب يعاني أحيانا بضعف عضلات بعد التمرين الشاب يطلب العلاج من دكتور مختص لتقوية الأعصاب الشاب يحتاج إلى فيتامينات لتقوية الأعصاب الوزن ٥٥ الطول ١٦٩	ضعف العضلات بعد التمرين العمر ٢٢ سنة الوزن ٥٥ كجم الطول ١٦٩ سم	الضعف والم في العضلات بعد التمرين
Peace be upon you. A 22-year-old young man goes to the gym daily but sometimes feels weakness and muscle pain after training. He needs a specialized doctor to advise on proper treatment to strengthen the nerves or vitamins for nerve strengthening. Weight 55, Height 169.	The young man sometimes suffers from muscle weakness after exercise. He seeks treatment from a specialist doctor to strengthen the nerves. He needs vitamins for nerve strengthening. Weight 55, Height 169.	Muscle weakness after exercise. Age 22 years. Weight 55 kg. Height 169 cm.	Weakness and muscle pain after exercise.

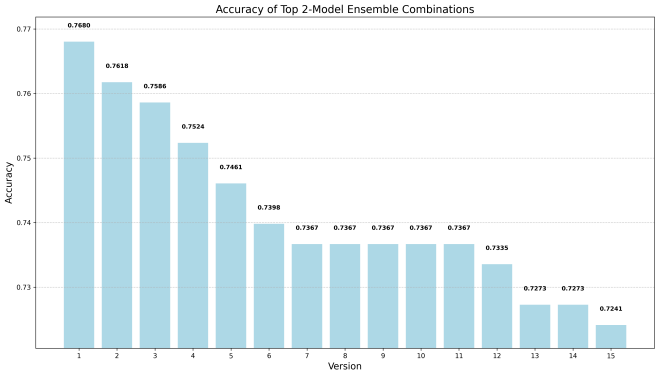

Fig. 3. Accuracy of Top 2-Model Ensemble Combinations. The best combination achieved an accuracy of 0.7680, showing limited performance improvement with only two models.

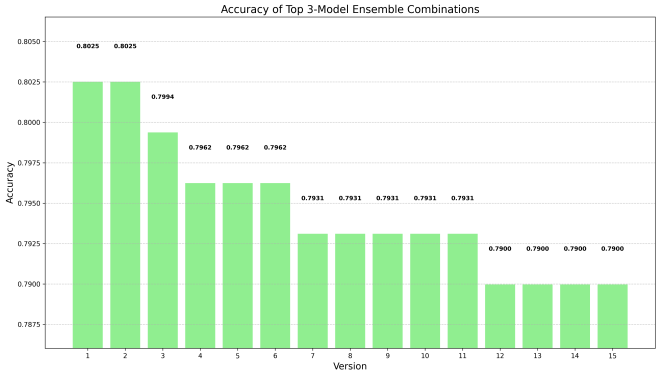

Fig. 4. Accuracy of Top 3-Model Ensemble Combinations. Accuracy improves significantly to 0.8025, indicating the benefits of model diversity.

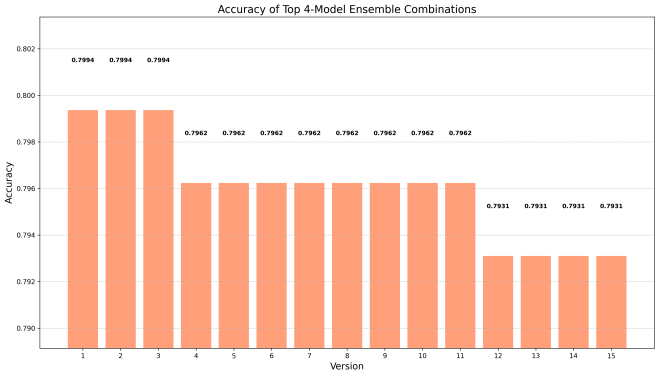

Fig. 5. Accuracy of Top 4-Model Ensemble Combinations. The best ensemble achieved 0.7994, with diminishing returns beyond three models.

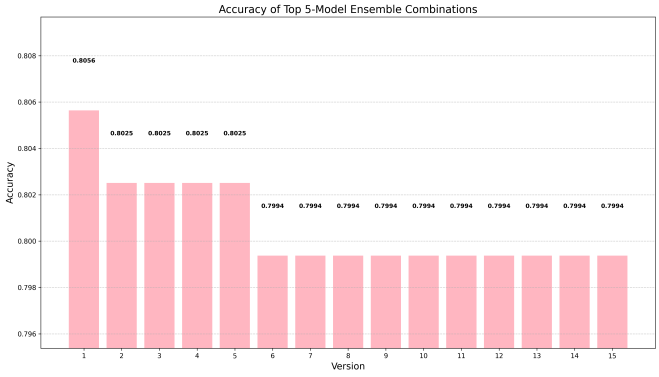

Fig. 6. Accuracy of Top 5-Model Ensemble Combinations. Reaches a peak of 0.8056, one of the best overall performances across all configurations.

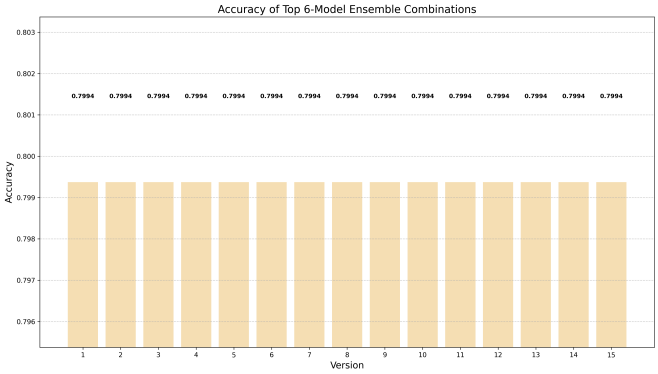

Fig. 7. Accuracy of Top 6-Model Ensemble Combinations. All combinations plateaued at 0.7994, suggesting ensemble saturation.

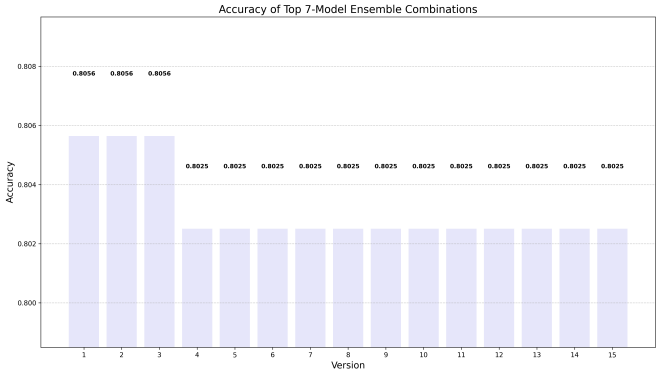

Fig. 8. Accuracy of Top 7-Model Ensemble Combinations. The highest accuracy of 0.8056 indicates effective scaling with selected models.

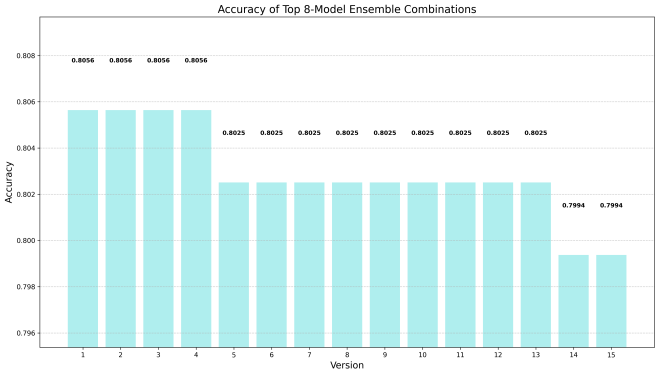

Fig. 9. Accuracy of Top 8-Model Ensemble Combinations. Matches best results at 0.8056, maintaining high performance.

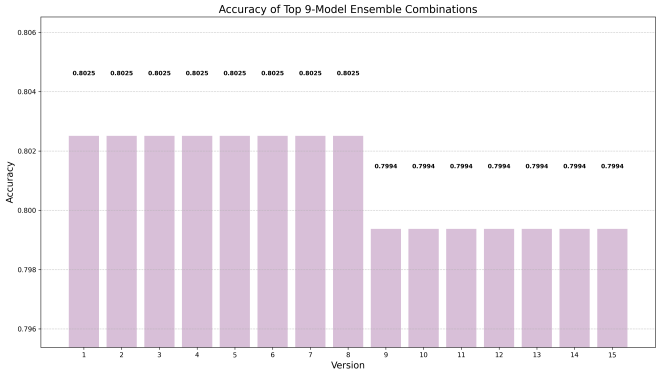

Fig. 10. Accuracy of Top 9-Model Ensemble Combinations. Peak accuracy is slightly lower at 0.8025, indicating diminishing benefits.

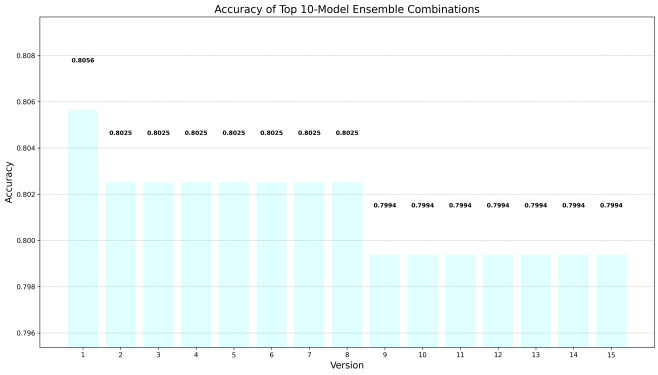

Fig. 11. Accuracy of Top 10-Model Ensemble Combinations. Matches the highest accuracy of 0.8056, showing that larger ensembles can still be optimal.

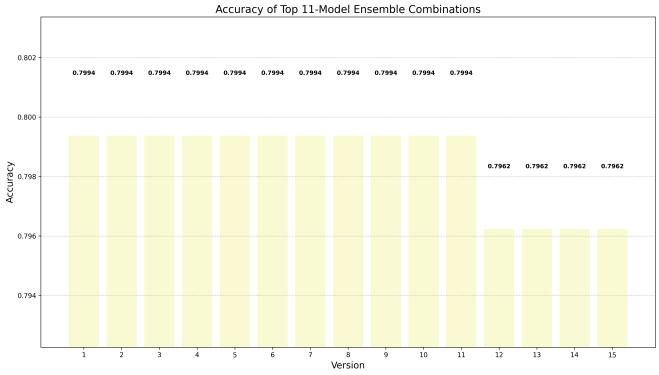

Fig. 12. Accuracy of Top 11-Model Ensemble Combinations. Maximum accuracy observed was 0.7994, with reduced variation across versions.

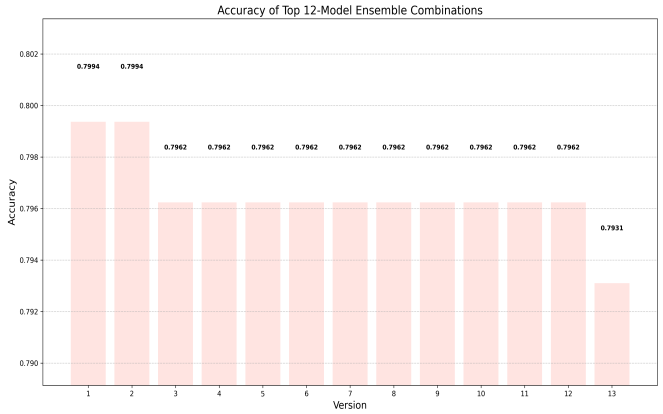

Fig. 13. Accuracy of Top 12-Model Ensemble Combinations. The highest recorded was 0.7994.

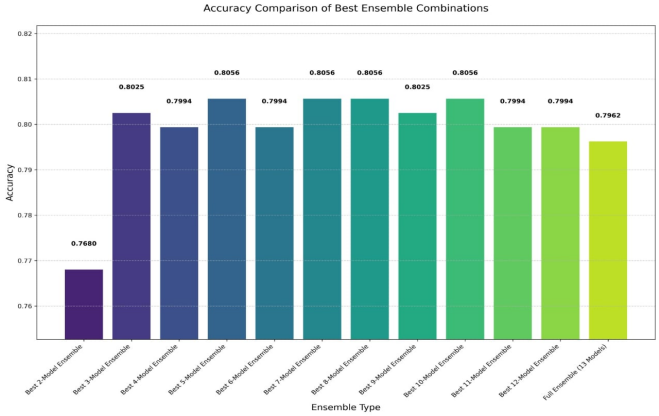

Fig. 14. Comparison of Best Ensemble Combinations Across Different Ensemble Sizes. The highest recorded accuracy was 0.8056, observed consistently across the 4, 6, 7, and 10-model ensembles.

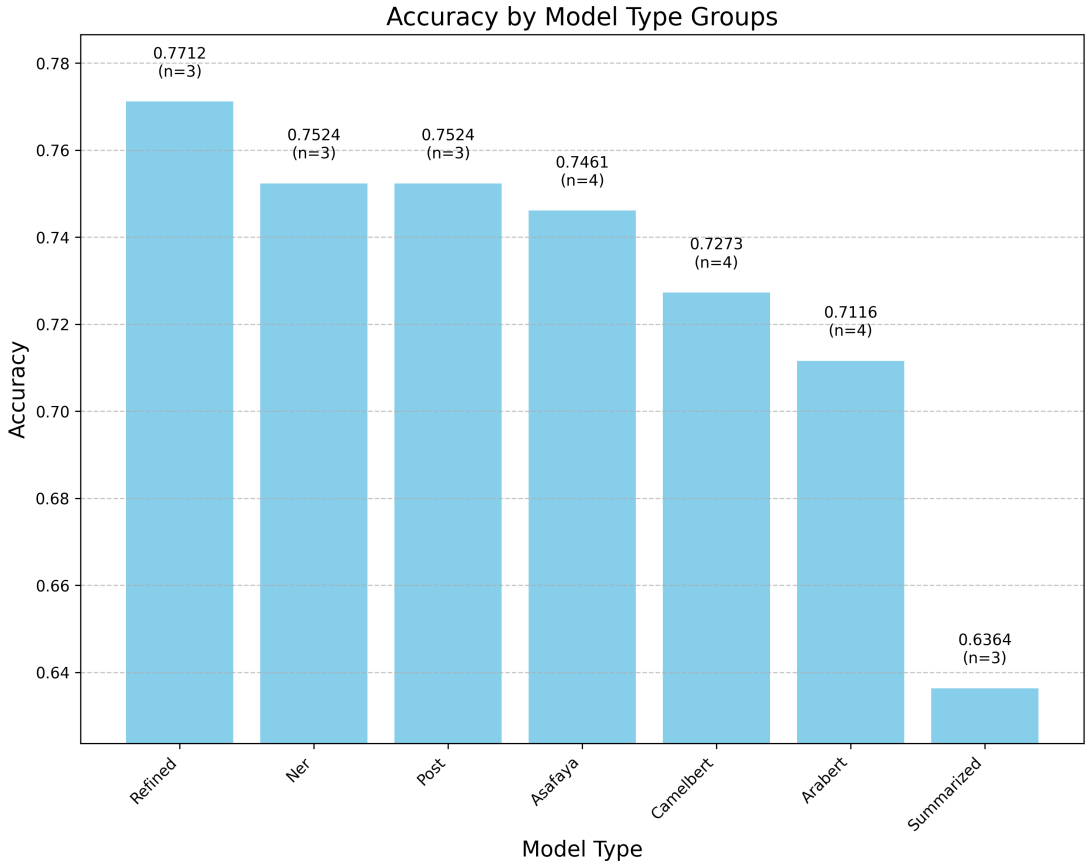

Fig. 15. Accuracy by Model Type Groups. The ensemble with refined models achieved the highest accuracy, while summarized models performed the worst.

TABLE II
PERFORMANCE COMPARISON BETWEEN INDIVIDUAL MODELS AND THE ENSEMBLE STRATEGY.

Model	Accuracy
CAMeLBERT	
CamelBERT_Post	70.53%
CamelBERT_Refined	75.55%
CamelBERT_NER	74.61%
CamelBERT_Summarized	64.89%
AraBERT	
AraBERT_Post	71.79%
AraBERT_Refined	72.41%
AraBERT_NER	68.97%
AraBERT_Summarized	65.94%
AsafayaBERT	
AsafayaBERT_Post	74.92%
AsafayaBERT_Refined	75.24%
AsafayaBERT_NER	73.67%
AsafayaBERT_Summarized	74.92%
Ensemble (Majority Voting)	**80.56%**

Overall, the results reveal that ensemble strategies improve disease classification performance when multiple diverse models are combined. However, the accuracy gains diminish beyond three models unless the ensemble is carefully curated. The best accuracy of 0.8056 was consistently observed across ensemble sizes of 5, 7, 8, and 10 models.

To further analyze ensemble performance, we conducted experiments by grouping models based on two criteria: (1) preprocessing augmentation type (Refined, NER, Post, Summarized) and (2) underlying model architecture (AraBERT, CAMeLBERT, AsafayaBERT). For each group, we created an ensemble by selecting all models belonging to that category and applied a majority voting strategy to generate final predictions.

As illustrated in Figure 15, the ensemble constructed using Refined models achieved the highest accuracy at 0.7712, followed closely by the NER and Post-based ensembles, both reaching 0.7524. In contrast, the ensemble using Summarized models performed the worst with an accuracy of 0.6364, indicating that excessive compression of medical content may hinder classification performance.

In terms of base models, the ensemble composed of AsafayaBERT variants outperformed those formed from CAMeLBERT and AraBERT models. The AraBERT-based ensemble recorded the lowest accuracy (0.7116) among the model-type groups. These results suggest that both the choice of preprocessing strategy and the underlying language model architecture play a critical role in ensemble-based classification accuracy.

This experiment highlights the benefit of majority voting ensembles that are strategically grouped, either by augmentation method or model family, for improving disease classification from Arabic medical texts.

VI. CONCLUSION

This work contributes in combining fine-tuned the Arabic language models for the purpose of disease classification in social telehealth applications. The methodology involving multi-layered preprocessing for text Refinement, Summarization, and Named Entity Recognition (NER) to extract key words of medical posts. The enriched datasets were used to fine-tune three notable Arabic language models: CAMeL-BERT, AraBERT, and Asafaya-BERT, combining them with majority voting ensembling technique to enhance the accuracya and the best performance achieved 80.56%. This approach improves performance and provides a new benchmark for its application in telehealth. Future studies can extend this framework to other languages or combining multiple models using an ensemble technique for severity assessment.

REFERENCES

[1] J. Chen and Y. Wang, "Social media use for health purposes: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, 2020.

[2] R. B. Correia, I. B. Wood, J. Bollen, and L. M. Rocha, "Mining social media data for biomedical signals and health-related behavior," *Annual Review of Biomedical Data Science*, vol. 3, no. 1, pp. 433–458, 2020.

[3] Y. Guo, A. Ovadje, M. A. Al-Garadi, and A. Sarker, "Evaluating large language models for health-related text classification tasks with public social media data," *Journal of the American Medical Informatics Association*, 2024.

[4] A. Magge *et al.*, "Overview of the sixth social media mining for health applications (smm4h) shared tasks at naacl 2021," Jan. 2021.

[5] O. H. Abdellaif, A. Nader, and A. Hamdi, "Lmrpa: Large language model-driven efficient robotic process automation for ocr," *arXiv preprint arXiv:2412.18063*, 2024.

[6] O. Abdellatif, A. Ayman, and A. Hamdi, "Lmv-rpa: Large model voting-based robotic process automation," *arXiv preprint arXiv:2412.17965*, 2024.

[7] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019.

[8] O. H. Abdellaif, A. N. Hassan, and A. Hamdi, "Erpa: Efficient rpa model integrating ocr and llms for intelligent document processing," in *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. IEEE, 2024, pp. 295–300.

[9] A. Hamdi, H. Kassab, M. Bahaa, and M. Mohamed, "Riro: Reshaping inputs, refining outputs unlocking the potential of large language models in data-scarce contexts," *arXiv preprint arXiv:2412.15254*, 2024.

[10] A. Hamdi, A. A. Mazrou, and M. Shaltout, "Llm-sem: A sentiment-based student engagement metric using llms for e-learning platforms," *arXiv preprint arXiv:2412.13765*, 2024.

[11] V. Radivchev and A. Nikolov, "Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles," 2019, pp. 691–695.

[12] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A fine-tuned bert-based transfer learning approach for text classification," 2022.

[13] O. Hamad, K. Shaban, and A. Hamdi, "Asem: Enhancing empathy in chatbot through attention-based sentiment and emotion modeling," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 1588–1601.

[14] S. Casola and A. Lavelli, "Fbk @ smm4h 2020: Roberta for detecting medications on twitter," 2020, pp. 101–103.

[15] B. Li and F. Rudzicz, "Torontocl at cmcl 2021 shared task: Roberta with multi-stage fine-tuning for eye-tracking prediction," pp. 4–9, 2021.
[16] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, "Analyzing elmo and distilbert on socio-political news classification," in *Proceedings of the Workshop on Automatic Extraction of Socio-political Events from News 2020*, USA, May 2020, pp. 9–18.
[17] L. Wassim, K. Mohamed, and A. Hamdi, "Llm-daas: Llm-driven drone-as-a-service operations from text user requests," *arXiv preprint arXiv:2412.11672*, 2024.
[18] M. de Bruyn, E. Lotfi, J. Buhmann, and W. Daelemans, "Bart for knowledge grounded conversations," in *CEUR Workshop Proceedings*, vol. 2666, 2020.
[19] X. Ou and H. Li, "Ynu @ dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis," 2020, pp. 4–9.
[20] ——, "Ynu_oxz @ haspeede 2 and ami: Xlm-roberta with ordered neurons lstm for classification task at evalita 2020," in *EVALITA Evaluation of NLP and Speech Tools for Italian*, vol. 2765, 2020, pp. 102–109.
[21] Y. Zhao and X. Tao, "Zyj123@dravidianlangtech-eacl2021: Offensive language identification based on xlm-roberta with dpcnn," in *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Parkville, Victoria: EACL, 2021, pp. 216–221.
[22] K. A. Das, A. Baruah, F. A. Barbhuiya, and K. Dey, "Ensemble of electra for profiling fake news spreaders," vol. September, 2020, pp. 22–25.
[23] O. Hosam Abdellaif, A. Nader, and A. Hamdi, "Lmrpa: Large language model-driven efficient robotic process automation for ocr," *arXiv e-prints*, pp. arXiv–2412, 2024.
[24] A. Mustar, S. Lamprier, and B. Piwowarski, "Using bert and bart for query suggestion," in *CEUR Workshop Proceedings*, vol. 2621, 2020.
[25] C. Popa and T. Rebedea, "Bart-tl: Weakly-supervised topic label generation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021, pp. 1418–1425.
[26] N. Niraula, S. Ayhan, B. Chidambaram, and D. Whyatt, "Multi-label classification with generative large language models," in *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, vol. 1, 2024.
[27] S. S. Alahmari, L. O. Hall, P. R. Mouton, and D. B. Goldgof, "Repeatability of fine-tuning large language models illustrated using qlora," *IEEE Access*, vol. 1, 2024.