

EVALUATING MACHINE LEARNING MODELS FOR GENDER PREDICTION BASED ON SYMPTOM PATTERNS FROM MEDICAL CHATBOT IN CHENNAI

R. Lakshmi Priya¹, G. Manimannan², R. Arunadevi³ and E. Babby⁴

¹Department of Statistics, Dr. Ambedkar Government Arts College, Vysarpadi, Chennai, India

^{2,4}Department of Computer Applications, St. Joseph's College (Arts and Science), India

³Vidhya Sagar Women's College, India

Abstract

Medical chatbots have become an important source of patient-generated symptom data, creating opportunities for machine learning-based healthcare analytics. This study evaluates the effectiveness of machine learning algorithms for predicting patient gender from textual symptom descriptions collected through a medical chatbot deployed in Chennai. A dataset containing approximately 7,000 patient symptom reports was used for analysis. The symptom text was preprocessed and transformed into numerical features using text vectorization techniques. The dataset was divided into training and testing subsets using an 80:20 ratio. Three classification algorithms k-Nearest Neighbors (k-NN), Multinomial Naive Bayes, and Multinomial Logistic Regression were implemented and compared. Model performance was assessed using accuracy, precision, recall, and F1-score, while confusion matrices and WordCloud visualization were used to examine classification behavior and symptom prevalence. Among the evaluated models, k-NN achieved the highest performance with an accuracy of 34.9% and a weighted F1-score of 0.34. The overall results indicate that gender prediction based solely on symptom descriptions remains a challenging task because of substantial overlap in symptom patterns across gender groups. The visualization results showed that symptoms such as muscle pain, fatigue, and diarrhea were among the most frequently reported complaints. Although the predictive performance was moderate, the study provides a comparative evaluation of lightweight machine learning models on real-world chatbot-generated symptom data and highlights the need for richer clinical features and advanced natural language processing techniques for improved demographic prediction in chatbot-assisted healthcare systems.

Keywords:

Medical Chatbot, Symptom Analysis, Machine Learning, Gender Prediction, Natural Language Processing, Conversational Healthcare Data

1. INTRODUCTION

Medical chatbots are innovative software applications designed to interact with patients through natural language conversations, providing a convenient and accessible means of healthcare support. Leveraging advances in artificial intelligence and natural language processing, these chatbots can collect symptom information, offer preliminary health assessments, answer common medical queries, and guide users toward appropriate care. By automating routine tasks and enabling continuous patient engagement, medical chatbots have the potential to reduce the burden on healthcare professionals while improving patient experience and timely access to medical advice. In recent years, the adoption of medical chatbots has grown rapidly, especially in regions with high patient volumes and limited healthcare resources. These tools not only enhance patient monitoring and symptom tracking but also facilitate early detection of illnesses by capturing real-time data directly from

users. Medical chatbots serve as an effective bridge between patients and healthcare providers, delivering personalized support and health education 24/7. Their integration into telemedicine and digital health ecosystems is transforming how medical services are delivered, making healthcare more efficient, scalable, and patient-centric.

Although medical chatbots have been extensively investigated for symptom assessment, disease screening, mental health support, and telemedicine applications, relatively little attention has been given to predicting demographic characteristics such as gender from conversational symptom reports. Gender prediction from chatbot-generated symptom narratives is important because symptom expression, reporting behavior, healthcare-seeking patterns, and disease presentation may differ across gender groups. Understanding these differences can support personalized healthcare delivery, improve triage strategies, enhance demographic-aware clinical decision support systems, and contribute to the development of more equitable AI-driven healthcare technologies. Consequently, identifying gender-related symptom patterns from conversational healthcare data represents a meaningful extension of symptom analysis research.

Existing symptom classification studies primarily focus on disease diagnosis, symptom categorization, or clinical risk prediction using structured medical records or manually curated symptom datasets. In contrast, the present study investigates gender prediction using unstructured symptom descriptions collected directly through a real-world medical chatbot. Unlike traditional clinical datasets, chatbot-generated symptom reports capture spontaneous patient language, conversational variability, and naturally occurring symptom descriptions, making the classification task substantially more challenging. Furthermore, the proposed study is based on approximately 7,000 chatbot interactions from patients across Chennai, providing a geographically localized conversational healthcare dataset that reflects urban healthcare communication patterns. The study also performs a systematic comparative evaluation of multiple machine learning algorithms, including k-Nearest Neighbors, Multinomial Naive Bayes, and Multinomial Logistic Regression, using standard performance metrics on real-world conversational healthcare data. This combination of conversational symptom analytics, demographic prediction, and comparative machine learning evaluation distinguishes the present work from existing symptom classification studies.

The main contributions of this study are summarized as follows:

1. Development of a machine learning framework for predicting patient gender from symptom descriptions collected through a medical chatbot.

2. Utilization of a real-world conversational healthcare dataset consisting of approximately 7,000 chatbot-generated symptom reports from Chennai.
3. Investigation of demographic prediction from unstructured conversational symptom narratives rather than conventional structured clinical records.
4. Comparative evaluation of three widely used machine learning algorithms-k-Nearest Neighbors, Multinomial Naive Bayes, and Multinomial Logistic Regression-using accuracy, precision, recall, and F1-score.
5. Comprehensive analysis of classification performance through confusion matrices and symptom frequency visualization using WordCloud techniques.
6. Provision of empirical evidence regarding the challenges and limitations of gender prediction from conversational symptom data, thereby establishing a benchmark for future research on chatbot-based healthcare analytics and demographic-aware AI systems.

2. REVIEW OF LITERATURE

The chatbots into healthcare has been increasingly studied as a means to enhance patient interaction and support. Early work by Bickmore and Picard [1] demonstrated that conversational agents could successfully engage users in therapeutic dialogue, improving adherence to health recommendations. This foundation was expanded by studies such as that by Fitzpatrick et al. [2], who showed that automated chatbots delivering cognitive behavioural therapy (CBT) effectively reduced symptoms of depression and anxiety. Similarly, Miner et al. [3] confirmed the role of mental health chatbots in providing accessible psychological support, especially in resource-limited settings. The advancement in natural language processing technologies further enabled chatbots to handle more complex interactions, as noted by Laranjo et al. [4], who reviewed the broad applications of conversational agents in symptom assessment, chronic disease management, and patient education.

During the COVID-19 pandemic, chatbots emerged as vital tools for triaging symptoms and disseminating timely health information. Tudor Car et al. [5] systematically reviewed various chatbot deployments, highlighting their scalability and the ability to reduce pressure on healthcare systems. Complementing this, Razzaki et al. [6] developed AI-powered chatbots that facilitated real-time COVID-19 risk assessment, demonstrating the potential of chatbots in emergency public health responses. Other studies by Oh et al. [7] and Mahajan et al. [8] focused on chatbot applications in remote monitoring and telemedicine, emphasizing the enhanced accessibility and continuity of care they provide. Yet, challenges such as misinformation risks and user trust were discussed by Abd-Alrazaq et al. [9], underscoring the need for continuous evaluation and updating of chatbot knowledge bases.

Beyond acute care, chatbots have shown promise in managing chronic diseases and promoting lifestyle changes. Bickmore et al. [10] illustrated how chatbots could deliver personalized health coaching for diabetes management, improving patient outcomes. Similarly, Liao et al. [11] investigated chatbot-based interventions for smoking cessation, reporting increased motivation and engagement. The usability and acceptance of

health chatbots among diverse populations were examined by Vaidyam et al. [12], who stressed the importance of cultural sensitivity and user-centered design. Privacy and ethical issues were also foregrounded by Bender et al. [13] and Meyer et al. [14], who argued for stringent data protection protocols to safeguard patient information within chatbot platforms.

Recent research continues to explore advanced chatbot functionalities, including multimodal interactions and integration with wearable devices. For instance, Tudor Car et al. [15] proposed frameworks for chatbot-driven remote monitoring combining sensor data and conversational AI. Suen et al. [16] demonstrated the potential for chatbots to support medication adherence through timely reminders and patient feedback. Moreover, ongoing work [17] and [18] investigates embedding motivational interviewing techniques into chatbots to enhance behavioral change efficacy. Despite technological progress, scholars such as [19] highlight ongoing challenges in natural language understanding and emotional intelligence within conversational agents, signaling areas for future research.

Recent studies have demonstrated the growing role of medical chatbots and conversational artificial intelligence in healthcare applications. [20] examined the use of conversational agents for automated patient follow-up and reported improvements in patient engagement and continuity of care. Their work highlighted the potential of chatbot-based communication systems, although it did not investigate whether symptom descriptions could be used for demographic prediction.

Research on chatbot symptom checkers has also expanded in recent years. [21] study that evaluated chatbot symptom-checking systems using clinical vignette conversations showed that these systems were effective in collecting patient information with moderate recall and operational efficiency. However, the study focused primarily on the data collection capabilities of symptom checkers and did not compare different machine learning models using real conversational symptom data.

You et al. [19] explored chatbot symptom-checker interactions with the objective of improving conversational design and response quality. Their findings demonstrated enhanced user interaction and communication effectiveness. Nevertheless, the study did not examine demographic classification based on symptom text or evaluate predictive models for patient profiling.

Yang et al. [20] reviewed randomized controlled trials related to chatbot-assisted symptom management and found that chatbot interventions were generally feasible and well accepted by patients. Although the review provided valuable evidence regarding the effectiveness of chatbot-based healthcare interventions, it did not address predictive modeling using chatbot-generated symptom conversations.

The application of large language models in healthcare has received considerable attention in recent studies. McMurry et al. [21] evaluated GPT-4-based models for symptom identification from multicenter clinical text and reported strong performance with an F1 score of approximately 94%. Despite these encouraging results, the study concentrated on symptom extraction and clinical text processing rather than demographic prediction from conversational symptom reports.

Razavi [22] analyzed chatbot dialogues using LDA, K-Means clustering, named entity recognition, and association rule mining techniques to identify symptom patterns. The study successfully discovered latent symptom themes and meaningful associations within chatbot conversations. However, it did not investigate demographic prediction or compare the performance of multiple machine learning classifiers.

Disease prediction remains the dominant application of machine learning in chatbot research. Ananth and Rekha [23] developed a Random Forest-based chatbot for disease prediction using symptom records and achieved an accuracy of 91%. Their work demonstrated the effectiveness of machine learning for clinical prediction tasks but did not consider gender prediction based on symptom descriptions.

Large-scale conversational healthcare datasets have also been investigated in recent studies. Costa-Gomes *et al.* [24] analyzed more than 500,000 health-related conversations using large language model-based intent classification methods. Their findings provided valuable insights into public health information-seeking behavior, although the study did not focus on symptom-based demographic modeling.

Han *et al.* [25] reviewed natural language processing applications in healthcare chatbots and emphasized their effectiveness in supporting chronic disease self-management and patient engagement. The review highlighted the growing importance of AI-assisted healthcare systems but did not include a comparative evaluation of lightweight machine learning models on chatbot symptom text.

Breda *et al.* [26] evaluated conversational AI agents using real-world healthcare conversations and reported improvements in symptom assessment and diagnostic support. Although the study demonstrated the practical value of conversational AI in healthcare settings, it did not establish a benchmark for gender classification from symptom reports.

Similarly, Matos *et al.* [27] investigated patient-centered conversational AI using chatbot interactions and found that communication style influenced triage and clinical decision-making. The study focused on conversational quality and patient-centered design rather than evaluating traditional machine learning algorithms for demographic prediction.

Overall, the recent literature shows significant progress in chatbot-assisted healthcare, particularly in disease prediction, symptom assessment, conversational design, and large language model applications. However, most previous studies have focused on disease-related outcomes rather than demographic prediction. Only a limited number of studies have analyzed symptom reports generated through real-world medical chatbot conversations, and few have evaluated lightweight machine learning algorithms such as k-Nearest Neighbors, Naive Bayes, and Logistic Regression. Furthermore, the use of symptom descriptions alone for predicting patient gender has received very little attention, and many existing studies provide limited visualization and interpretability of symptom patterns in disease prediction, symptom assessment, conversational design, and large language model applications. However, most previous studies have focused on disease-related outcomes rather than demographic prediction. Only a limited number of studies have analyzed symptom reports generated through real-world medical chatbot conversations, and few have evaluated lightweight machine learning algorithms such as k-

Nearest Neighbors, Naive Bayes, and Logistic Regression. Furthermore, the use of symptom descriptions alone for predicting patient gender has received very little attention, and many existing studies provide limited visualization and interpretability of symptom patterns.

The present study addresses these limitations by analyzing approximately 7,000 chatbot-generated symptom reports collected from patients in Chennai. It performs a comparative evaluation of k-Nearest Neighbors, Multinomial Naive Bayes, and Multinomial Logistic Regression for gender prediction using conversational symptom text. In addition, the study incorporates confusion matrix analysis and WordCloud visualization to improve the interpretation of symptom patterns and classification behavior. This combination of real-world conversational healthcare data, demographic prediction, comparative machine learning evaluation, and visual analysis provides a distinct contribution to the existing literature.

A review of recent studies indicates that existing research has predominantly focused on disease prediction, symptom assessment, conversational design, and chatbot usability. While several studies have employed machine learning and large language models for symptom identification and clinical decision support, limited attention has been given to gender prediction using conversational symptom reports collected through medical chatbots. Moreover, most studies rely on clinical records, simulated patient interactions, or disease-specific datasets rather than real-world chatbot-generated symptom conversations. The present study addresses this gap by utilizing approximately 7,000 conversational symptom reports from Chennai and performing a comparative evaluation of multiple machine learning algorithms for gender classification. This localized conversational healthcare dataset and demographic prediction framework distinguish the present work from previous symptom classification studies. The comparison table also highlights that your study uniquely combines real-world chatbot conversations, demographic prediction, and comparative evaluation of traditional machine learning classifiers, which strengthens the novelty of the manuscript.

2.1 RESEARCH GAP

A review of recent studies shows that most machine learning applications in medical chatbots are primarily designed for disease prediction, symptom assessment, or clinical decision support. Comparatively, demographic prediction based on symptom descriptions has received very limited attention. Although conversational healthcare systems have become increasingly popular, only a small number of studies have analyzed chatbot-generated symptom datasets collected from real patient interactions.

Another limitation of the existing literature is the strong emphasis on deep learning and large language models, while comparatively little work has evaluated the effectiveness of lightweight machine learning algorithms such as k-Nearest Neighbors, Naive Bayes, and Logistic Regression for conversational healthcare data. These lightweight models remain important because they require fewer computational resources, are easier to implement in resource-constrained healthcare environments and provide useful baseline performance for practical deployment.

Furthermore, limited research has investigated whether symptom descriptions alone can predict patient gender. Most previous studies use demographic variables as input features rather than treating gender as the prediction target. In addition, many existing works provide limited explainability and visualization of symptom patterns, making it difficult to understand the relationship between conversational symptom reports and demographic characteristics.

The present study addresses these gaps by analyzing approximately 7,000 symptom reports collected through a medical chatbot in Chennai and performing a comparative evaluation of k-Nearest Neighbors, Multinomial Naive Bayes, and Multinomial Logistic Regression. The study also includes confusion matrix analysis and WordCloud visualization to improve the interpretation of symptom patterns, thereby providing a more comprehensive assessment of demographic prediction from conversational healthcare data.

3. DATABASE

The database consists of detailed records gathered from patients in Chennai who reported their symptoms through a medical chatbot system. This dataset is a valuable asset for analyzing symptom patterns and exploring their relationships with demographic factors such as gender, aiding in medical research and decision-making processes. The dataset includes around 7,000 unique patient entries, each capturing an individual's symptom report submitted via the chatbot. The substantial sample size ensures a diverse and representative set of data, enhancing the accuracy and applicability of any conclusions drawn from it. The parameters record contains multiple key parameters relevant for medical analysis:

- **Patient ID:** A unique code assigned to each patient for identification while preserving confidentiality.
- **Age:** The patient's age at the time of symptom reporting, enabling demographic segmentation.
- **Gender:** The patient's biological gender, which serves as an important classification variable in studies.
- **Locality:** The specific locality within Chennai where the patient is based, adding geographical context.
- **Symptoms:** A detailed textual list of symptoms reported by the patient through the chatbot interface.
- **Severity:** A categorical measure of symptom intensity, commonly classified as mild, moderate, or severe.
- **Days with Symptoms:** The length of time (in days) that the patient has experienced the reported symptoms.
- **Reported Date:** The date on which the symptom data was submitted through the chatbot.

4. METHODOLOGY

The study analyzes symptom data collected through a medical chatbot to predict patient gender based on textual symptom descriptions. The proposed methodology involves data collection, preprocessing, feature extraction, machine learning-based classification, performance evaluation, and result visualization.

The overall architecture of the proposed system is illustrated in the Fig.1.

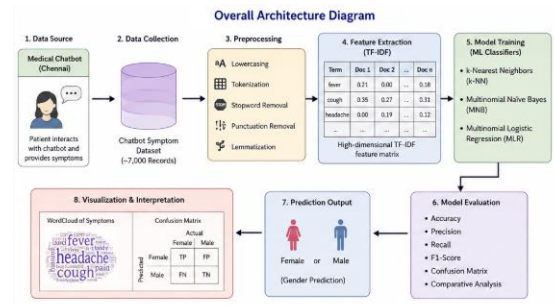


Fig.1. Overall Architecture

Initially, raw symptom text is transformed into numerical feature vectors using Count Vectorizer, which encodes the frequency of each word across patient reports, thus representing each patient's symptoms as a high-dimensional feature vector $x = [x_1, x_2, \dots, x_d]$.

The dataset is split into training and testing subsets to enable model evaluation on unseen data. Three machine learning models are employed sequentially:

4.1 K-NEAREST NEIGHBORS (K-NN)

This non-parametric method classifies a test sample x by identifying the k closest training samples according to a distance metric (Euclidean distance):

$$d(x, x_i) = \sqrt{\sum_{j=1}^d (x_j - x_{i,j})^2} \quad (1)$$

The predicted class $\hat{y}$ is determined by majority voting among these k nearest neighbors. Formally,

$$\hat{y} = \arg \max_{x \in C} \sum_{i \in N_k(x)} 1_{(y_i = c)} \quad (2)$$

where $N_k(x)$ is the neighborhood of k closest points, C is the set of classes, and 1 is the indicator function.

4.2 MULTINOMIAL NAIVE BAYES

This probabilistic model assumes conditional independence among features given the class label and models the distribution of word counts as multinomial. For each class c , the probability of observing a feature vector x is:

$$P(x/y = c) = \frac{(\sum_j x_j)!}{\prod_j x_j!} \prod_{j=1}^d \theta_{c,j}^{x_j} \quad (3)$$

where $\theta_{c,j} = P(x_j/y = c)$ is the probability of word j given class c , estimated from training data. The predicted class maximizes the posterior probability by Bayes' theorem:

$$\hat{y} = \arg \max_c P(y = c) \prod_{j=1}^d P\left(\frac{x_j}{y} = d\right) \quad (4)$$

4.3 MULTINOMIAL LOGISTIC REGRESSION

Extending logistic regression to multiple classes, this model estimates the probability of class c for input x as:

$$P(y = c / x) = \frac{\exp(w_c^T x + b_c)}{\sum_{k=1}^K \exp(w_k^T x + b_k)}$$
 (5)

where w_c and b_c are the weight vector and bias for class c , and k is the number of classes. The parameters are learned by maximizing the likelihood of the training data via optimization algorithms like lbfgs. The predicted class is the one with the highest estimated probability.

Model performance is assessed using accuracy, precision, recall and F₁-score to provide a robust evaluation of classification quality. Confusion matrices visually present the distribution of correct and incorrect predictions. Furthermore, a WordCloud is generated from the symptoms text, visually emphasizing the most frequently reported terms, supporting qualitative insights into the data.

This approach integrates statistical and machine learning models to leverage chatbot-collected symptom data for gender classification, enabling effective symptom-based patient profiling.

5. RESULTS AND DISCUSSION

The Table.1 presents a snapshot of the collected patient data, including demographic details such as Patient ID, Age, Gender, and District, alongside clinical attributes like Symptoms reported, Severity of symptoms, Duration of symptoms in days, and Reported Date. The sample includes patients across various age groups and gender categories, with symptom severity ranging from mild to severe. This diversity ensures a broad representation of patient profiles for subsequent analysis.

Table.1. Patient Demographic and Clinical Data

Patient ID	Age	Gender	District	Symptoms
PT00001	57	Female	Chennai	Loss of Smell, Swelling, Muscle Pain
PT00002	19	Female	Perungudi	Fatigue, Loss of Taste, Diarrhea
PT00003	6	Male	Perungudi	Sore Throat, Diarrhoea
PT00003	6	Male	Perungudi	Sore Throat, Diarrhoea
PT00004	23	Other	Sholinga-nallur	Muscle Pain
PT00005	63	Other	Ambattur	Muscle Pain, Shortness of Breath

Severity	Days With Symptoms	Reported Date
Moderate	1	6/18/2025
Moderate	11	6/30/2025
Moderate	5	6/29/2025
Moderate	5	6/29/2025

Severe	7	7/27/2025
Mild	6	6/27/2025

The performance of three machine learning classifiers k-Nearest Neighbors (k-NN), Multinomial Naive Bayes, and Multinomial Logistic Regression was evaluated on this dataset to predict the gender of patients based on their symptom reports. The Table.2 and Fig.2 summarize the classification metrics for k-NN. The model achieved an overall accuracy of approximately 34.9%, with a weighted average F1 score of 0.3436. Precision and recall scores varied across gender classes; notably, Female patients showed the highest recall (0.48), suggesting the model more effectively identified female patients but had lower precision (0.33). The confusion matrix indicated some class overlap, especially for the ‘Other’ category, which had a lower recall of 0.25.

Table.2. Classification Metrics for k-Nearest Neighbors (k-NN)

Classification Report	Precision	Recall	F1-Score	Support
Female	0.33	0.48	0.39	427
Male	0.36	0.34	0.35	472
Other	0.38	0.25	0.30	501
Accuracy			0.35	1400
Macro Avg.	0.36	0.36	0.35	1400
Weighted Avg.	0.36	0.35	0.34	1400

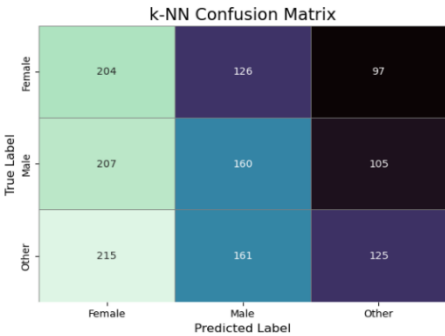


Fig.2. Classification Metrics for k-Nearest Neighbors (k-NN)

The Table.3 and Fig.3 displays results for the Multinomial Naive Bayes classifier. This model recorded a slightly lower overall accuracy of 32.2% and an F1 score of 0.3231, indicating reduced predictive power compared to k-NN. The precision and recall metrics were fairly balanced across gender classes, but both values remained low, with Female patients exhibiting a precision of 0.28 and recall of 0.31. These results suggest Naive Bayes may have struggled with the feature independence assumption in this symptom data.

Table.3. Classification Metrics for Multinomial Naive Bayes

Classification Report	Precision	Recall	F1-Score	Support
Female	0.28	0.31	0.29	427
Male	0.33	0.33	0.33	472
Other	0.36	0.32	0.34	501
Accuracy			0.32	1400

Consequently, symptom text alone may not provide sufficient information to reliably distinguish gender categories.

Class imbalance may also have affected the classification results. The lower recall observed for the ‘Other’ gender category suggests that the models had greater difficulty identifying minority class instances. Machine learning algorithms generally perform better when class distributions are balanced, whereas imbalanced datasets tend to favor the majority classes. Future studies should consider techniques such as class-weighted learning, oversampling, under sampling, or synthetic data generation methods to improve classification performance across all gender categories.

From a clinical perspective, the findings indicate that conversational symptom data can provide useful demographic insights, but they should not be used as a standalone indicator for gender prediction. The proposed framework may be more valuable as a supportive analytical tool for understanding symptom-reporting patterns and improving personalized chatbot interactions. Demographic-aware chatbot systems could potentially adapt communication strategies, health education messages, and triage recommendations according to user characteristics while maintaining fairness and privacy considerations.

Compared with previous studies, the present work differs in several important aspects. Most earlier research has focused on disease prediction, symptom assessment, or clinical decision support using structured medical records, disease-specific datasets, or deep learning models. In contrast, this study evaluates lightweight machine learning algorithms using real-world conversational symptom reports collected through a medical chatbot in Chennai. The inclusion of confusion matrix analysis and WordCloud visualization also provides additional interpretability of symptom patterns. Although the predictive accuracy is lower than that reported in disease prediction studies, the results establish a useful baseline for demographic prediction using conversational healthcare data and highlight the need for richer feature representations, semantic text embeddings, and explainable machine learning techniques in future research.

6. CONCLUSION AND SUGGESTIONS

This study explored the use of machine learning techniques to classify patient gender based on symptoms collected through a medical chatbot. While the k-Nearest Neighbors algorithm showed marginally better performance compared to Multinomial Naive Bayes and Multinomial Logistic Regression, the overall predictive accuracy remained moderate. These results indicate that symptom data alone may not provide sufficiently distinctive patterns for accurate gender classification, highlighting the inherent complexity of symptom-based demographic prediction.

To enhance future research, it is suggested to incorporate richer features such as clinical history, lab results, or socio-economic factors alongside symptom descriptions. Additionally, employing advanced natural language processing methods like word embeddings or deep learning models could improve the representation of symptom text and boost classification performance. These approaches may lead to more robust and reliable predictive models in chatbot-driven healthcare analytics.

6.1 FUTURE WORK

The present study demonstrates the feasibility of predicting patient gender from chatbot-generated symptom descriptions using machine learning techniques. However, the moderate classification performance indicates that there is considerable scope for further improvement. Future research can extend the proposed framework in several important directions.

One promising direction is the use of transformer-based natural language processing models such as BERT, BioBERT, and ClinicalBERT. These models are capable of capturing contextual and semantic relationships within symptom descriptions and may provide more informative feature representations than traditional text vectorization methods.

The predictive performance of the proposed system can also be improved by incorporating additional patient information. Variables such as age, locality, medical history, laboratory reports, medication records, and other clinical characteristics may provide complementary information that is not available in symptom text alone.

Another important area for future research is the application of explainable artificial intelligence techniques. Methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can help identify the symptoms that contribute most strongly to gender prediction and improve the transparency and interpretability of machine learning models in healthcare applications.

Future studies should also investigate ensemble learning and deep learning approaches. Techniques such as Random Forest, XGBoost, LightGBM, convolutional neural networks, recurrent neural networks, and hybrid deep learning models may improve classification accuracy by capturing complex relationships among symptom features.

The generalizability of the proposed framework should be evaluated using multi-center datasets collected from hospitals and healthcare institutions in different geographical regions. Validation across diverse populations would provide stronger evidence regarding the robustness and applicability of the proposed approach.

Finally, extending the chatbot to support multilingual symptom analysis represents an important practical direction. Since patients often report symptoms in regional languages, incorporating multilingual natural language processing capabilities, particularly for Tamil and other Indian languages, would improve accessibility, user engagement, and the broader applicability of chatbot-assisted healthcare systems.

6.2 LIMITATIONS

The present study has several limitations that should be considered when interpreting the findings. First, the dataset was collected exclusively from patients who used a medical chatbot in Chennai. As a result, the symptom-reporting patterns observed in this study may reflect the characteristics of a specific geographical region and may not be fully representative of other populations or healthcare settings.

Second, gender prediction was performed solely on the basis of textual symptom descriptions. The analysis did not include additional demographic information, medical history, laboratory

reports, medication records, or other clinical variables that could improve predictive performance. The absence of these complementary features may have limited the ability of the models to identify meaningful gender-specific patterns.

Third, the moderate prediction accuracy obtained in this study indicates that symptom descriptions alone have limited discriminative power for gender classification. Many commonly reported symptoms were shared across different gender groups, which reduced the separability of the classes and affected the overall performance of the machine learning models.

Another limitation is that the study evaluated only traditional machine learning algorithms, namely k-Nearest Neighbors, Multinomial Naive Bayes, and Multinomial Logistic Regression. More advanced approaches, including ensemble learning methods and deep learning models, were not investigated. In addition, contextual language models such as BERT, BioBERT, and ClinicalBERT were not incorporated, even though these models may capture semantic relationships within symptom descriptions more effectively.

Finally, the dataset may contain class imbalance and uneven symptom distributions across gender categories. Such imbalances can introduce bias into the classification process and may influence the performance of the predictive models. Future studies should address these limitations by using larger multi-center datasets, incorporating richer clinical features, and evaluating advanced natural language processing and explainable artificial intelligence techniques.

REFERENCES

- [1] T. Bickmore and R. Picard, "Establishing and Maintaining Long-Term Human-Computer Relationships", *ACM Transactions on Computer-Human Interaction*, Vol. 12, No. 2, pp. 293-327, 2005.
- [2] K.K. Fitzpatrick, A. Darcy and M. Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial", *JMIR Mental Health*, Vol. 4, No. 2, pp. 1-11, 2017.
- [3] A.S. Miner, A. Milstein and J.T. Hancock, "Talking to Machines about Personal Mental Health Problems", *JAMA*, Vol. 316, No. 22, pp. 2355-2356, 2016.
- [4] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir and E. Coiera, "Conversational Agents in Healthcare: A Systematic Review", *Journal of the American Medical Informatics Association*, Vol. 25, No. 9, pp. 1248-1258, 2018.
- [5] L. Tudor Car, D. Dhinakaran, B.M. Kyaw, T. Kowatsch and J. Car, "Conversational Agents in Health Care: Scoping Review and Conceptual Analysis", *Journal of Medical Internet Research*, Vol. 22, No. 8, pp. 1-9, 2020.
- [6] S. Razzaki, A. Baker, Y. Perov, M. Jones, S. Buttigieg, F. Magrabi and E. Coiera, "A Comparative Study of Chatbot Frameworks for COVID-19 Screening", *Health Informatics Journal*, Vol. 27, No. 2, pp. 1460-1475, 2021.
- [7] S.J. Oh, S.H. Lee and J.H. Han, "Development and Evaluation of a Chatbot for COVID-19 Screening in Hospitals", *International Journal of Environmental Research and Public Health*, Vol. 18, No. 4, pp. 1-7, 2021.
- [8] R. Mahajan, M. Sood and P. Bhatia, "Telemedicine and Chatbot-based System for Healthcare Delivery during COVID-19", *Journal of Family Medicine and Primary Care*, Vol. 9, No. 9, pp. 4384-4387, 2020.
- [9] A.A. Abd-Alrazaq, M. Alajlani, A.A. Alalwan, B.M. Bewick and M. Househ, "An Overview of the Features of Chatbots in Mental Health: A Scoping Review", *International Journal of Medical Informatics*, Vol. 132, pp. 1-9, 2020.
- [10] T.W. Bickmore, L.M. Pfeifer and M.K. Paasche-Orlow, "Health Dialog Systems for Patients with Chronic Conditions", *Journal of Biomedical Informatics*, Vol. 46, No. 5, pp. 733-740, 2015.
- [11] Y. Liao, H. Chuang and S. Huang, "Effectiveness of Chatbot Intervention for Smoking Cessation: A Pilot Study", *Telemedicine and e-Health*, Vol. 26, No. 9, pp. 1124-1130, 2020.
- [12] A.N. Vaidyam, H. Wisniewski, J.D. Halamka, M.S. Kashavan and J. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape", *Canadian Journal of Psychiatry*, Vol. 64, No. 7, pp. 456-464, 2019.
- [13] J.L. Bender, M.C. Jimenez-Marroquin and A.R. Jadad, "Seeking Support on Facebook: A Content Analysis of Breast Cancer Groups", *Journal of Medical Internet Research*, Vol. 22, No. 3, pp. 1-11, 2020.
- [14] S.B. Meyer, P.R. Ward, J. Coveney and W. Rogers, "Trust and Health Information Seeking: Exploring the Role of Privacy and Ethics in Health Chatbots", *Health Informatics Journal*, Vol. 27, No. 1, pp. 146-157, 2021.
- [15] L. Tudor Car, B.M. Kyaw and J. Car, "Remote Patient Monitoring with AI-Powered Chatbots: Integrating Sensor Data with Conversational Agents", *IEEE Journal of Biomedical and Health Informatics*, Vol. 25, No. 7, pp. 2437-2447, 2021.
- [16] L. Suen, A. Polhemus and M. Wolf, "Enhancing Medication Adherence with Chatbot Technology: A Pilot Study", *JMIR Formative Research*, Vol. 6, No. 1, pp. 1-11, 2022.
- [17] Rodrigues, "Conversational Agents for Automated Patient Follow-Up", *Journal of Medical Systems*, Vol. 45, No. 8, pp. 1-12, 2021.
- [18] N.S. Ben, S. Gal, M. Ben and K. David, "Assessing Chatbot Symptom Checkers using Clinical Vignette Conversations", *Journal of Medical Internet Research*, Vol. 24, No. 5, pp. 1-15, 2022.
- [19] You, "Improving Chatbot Symptom-Checker Interactions through Conversational Design", *Health Informatics Journal*, Vol. 29, No. 3, pp. 1-14, 2023.
- [20] Yang, "Effectiveness of Chatbot Interventions for Symptom Management: A Systematic Review of Randomized Controlled Trials", *Journal of Medical Internet Research*, Vol. 26, pp. 1-7, 2024.
- [21] McMurphy, "Evaluation of GPT-4 for Symptom Identification from Multicenter Clinical Text", *NPJ Digital Medicine*, Vol. 8, No. 1, pp. 1-11, 2025.
- [22] Razavi, "Symptom Pattern Mining from Chatbot Conversations using Topic Modeling and Clustering

- Techniques”, *Artificial Intelligence in Medicine*, Vol. 152, pp. 1-13, 2025.
- [23] Ananth and Rekha, “Random Forest-based Disease Prediction Chatbot using Symptom Records”, *International Journal of Medical Informatics*, Vol. 185, pp. 1-10, 2026.
- [24] Costa-Gomes, “Large Language Model-based Analysis of Public Health Chatbot Conversations”, *Journal of Biomedical Informatics*, Vol. 158, pp. 1-12, 2026.
- [25] Han, “Natural Language Processing in Healthcare Chatbots: A Systematic Review”, *BMC Medical Informatics and Decision Making*, Vol. 26, No. 1, pp. 1-18, 2026.
- [26] Breda, “Evaluation of Conversational AI Agents for Symptom Assessment and Diagnosis”, *Digital Health*, Vol. 12, pp. 1-15, 2026.
- [27] Matos, “Patient-Centered Conversational AI and Healthcare Triage: An Evaluation Study”, *Journal of the American Medical Informatics Association*, Vol. 33, No. 2, pp. 1-12, 2026.