International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03806
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

CLINICAL DIAGONSIS DECISION MAKING AND DRUG USAGE REGULATIONS

Dr. A. Suresh Rao [1], Waghmare Manisha [2], Suroju Gowtham [3], Sai Reddy Srivalli [4], Shanigarapu Srinivas [5].

[1] Head & Professor, Department of Computer and Science Engineering, TKR College of Engineering and Technology.

[2,3,4,5]UG Scholars, Department of Computer and Science Engineering, TKR College of Engineering and Technology, Medbowli, Meerpet.

ABSTRACT:

21st century, the emergence of new diseases with overlapping and additional symptoms has made disease prediction more complex. Early diagnosis plays a crucial role in reducing fatalities by enabling timely treatment. The existing systems rely on traditional machine learning models like Random Forest and Naïve Bayes for disease prediction but lack the efficiency and accuracy needed for precise diagnosis. This project proposes a Boosting-based disease prediction model using XGBoost and AdaBoost, along with comparative analysis against Light GBM, Extra Trees, K-Nearest Neighbors (KNN), Naïve Bayes, and Random Forest. The dataset consists of 1,200 datapoints, covering 24 diseases, with symptom descriptions in natural language. By leveraging Boosting techniques, the system enhances prediction accuracy through an ensemble approach. The system is implemented using Flask, providing an interactive web-based interface where users can input symptoms in textual format. The model then predicts the most probable disease and provides drug recommendations, precautionary measures, and dietary suggestions. The proposed approach demonstrates improved performance over traditional models, ensuring a more reliable disease diagnosis and prevention system.

KEYWORDS: Disease Prediction, Machine Learning, Boosting Algorithms, XGBoost, AdaBoost, Flask, Natural Language Processing, Light GBM , Extra Trees, Random Forest, Naïve Bayes, KNN, Symptom-Based Diagnosis

I. INTRODUCTION:

The advent of the 21st century has brought with it a rapid proliferation of new and emerging diseases, many of which exhibit similar or overlapping symptoms. This presents a substantial challenge to healthcare professionals, as timely and accurate diagnosis becomes increasingly difficult in the presence of ambiguous or co-occurring symptomatology. The necessity for efficient and intelligent disease prediction systems is more crucial than ever, with early diagnosis serving as a key determinant in effective treatment and mortality reduction.

Traditional machine learning models, such as Random Forest and Naïve Bayes, have been employed in diagnostic systems and have demonstrated foundational utility. However, these models often fail to process natural language-based symptom descriptions effectively, resulting in diminished predictive accuracy. To address these limitations, this work proposes a system based on ensemble learning, specifically leveraging boosting algorithms—XGBoost and AdaBoost—for disease classification. These algorithms offer iterative learning and error correction capabilities, enabling them to capture complex symptom interactions and improve classification outcomes.

The proposed system integrates these machine learning techniques into a Flask-based web application that accepts symptom input in plain text. The application subsequently processes the input, predicts the most probable disease, and provides users with tailored recommendations, including medication, dietary advice, and precautionary measures. This comprehensive approach enhances the precision, usability, and applicability of automated disease diagnostics in real-world scenarios.

II. LITERATURE SURVEY

The Numerous researchers have explored the application of machine learning in disease prediction and drug recommendation systems. Kumar et al. [1] introduced a hybrid system combining multiple classifiers with sentiment analysis for drug recommendation. While the model showed promise, it lacked robustness in handling rare diseases and relied heavily on data quality.

Singh and Kaur [2] proposed a voice-enabled diagnostic system utilizing Naïve Bayes and Decision Trees. However, the model's accuracy diminished in noisy environments and was constrained by limited disease coverage. Ramesh et al. [3] applied a fine- tuned AdaBoost classifier to predict chronic diseases such as heart and kidney disorders, but scalability and bias in datasets posed challenges.

Other notable contributions include Zhang et al.'s [4] use of deep learning and BERT for natural language-based disease prediction, and Gupta et al.'s [5] sentiment-enhanced drug recommendation engine. Despite improvements in predictive power, these systems often lacked interpretability and required high-quality, structured datasets.

This study builds on the limitations identified in the literature by incorporating boosting algorithms with multi-hot symptom encoding and synthetic data balancing techniques, ultimately delivering improved prediction accuracy and comprehensive healthcare recommendations.

III. METHODLOGY

The proposed methodology consists of several critical phases, including data preprocessing, feature engineering, class balancing, model training, and deployment via a web-based interface.

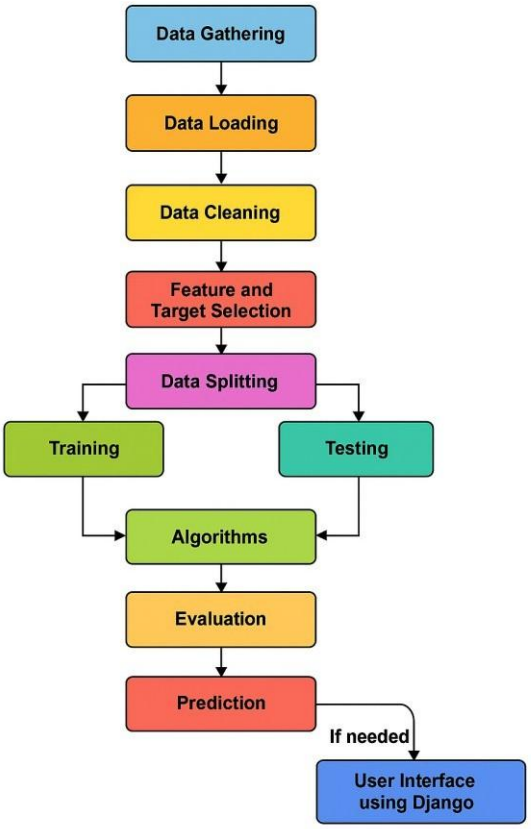

Fig 1: Block Diagram

A. Data Preprocessing

Missing values in both numerical and categorical data are imputed using mean and mode values respectively. Symptom text is

stripped of whitespace and standardized to ensure uniform encoding.

B. Feature Reduction and Encoding

Symptom descriptions are converted into binary vectors using a multi-hot encoding technique. This allows for efficient processing by machine learning models, preserving the presence or absence of specific symptoms across all records.

C. Class Balancing with SMOTE

To counteract class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. This technique synthesizes new data points for minority classes, ensuring balanced representation across disease categories.

D. Model Training

The primary models used are XGBoost and AdaBoost, with additional benchmarking conducted using Random Forest, K-Nearest Neighbors (KNN), LightGBM, Extra Trees, and Naïve Bayes classifiers. The dataset is split into training and test sets, and models are evaluated using F1-score, precision, and recall metrics.

E. System Deployment

The trained models are integrated into a Flask-based web application. Users input symptom descriptions via the web interface, and the system returns the most likely disease along with drug recommendations, dietary suggestions, and precautionary measures.

IV. FLOW CHART

The proposed clinical diagnosis and drug recommendation system is designed to streamline and enhance the accuracy of disease detection based on user-submitted symptoms.

The process begins when the user enters symptoms through a web-based interface. These inputs, often in free-text format, undergo preprocessing operations, including normalization, tokenization, and multi-hot encoding, to transform them into structured, machine-readable feature vectors.

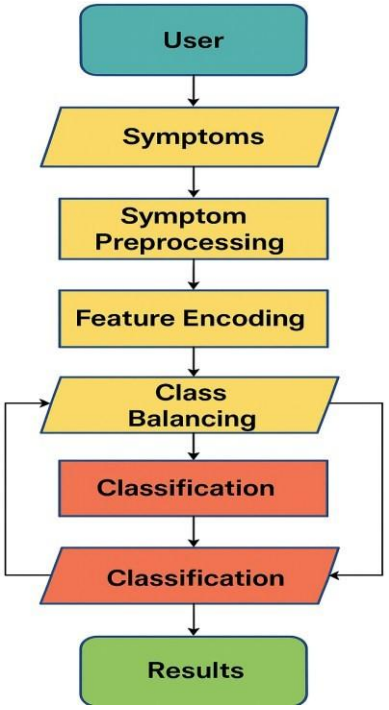

Fig 2: DATA FLOW DIAGRAM

Following this, the system applies class balancing using SMOTE (Synthetic Minority Over-sampling Technique) to address any disparities in disease representation within the dataset. The balanced and encoded data is then passed to an ensemble-based machine learning pipeline, where multiple classification algorithms—including XGBoost and AdaBoost—are evaluated. These boosting techniques iteratively correct prediction errors, leading to a higher diagnostic precision compared to traditional models such as Naïve Bayes or Random Forest. Upon successful classification, the system identifies the most probable disease and retrieves a curated set of clinical recommendations. These include

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03806
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

suitable drug options, dietary guidelines, and preventive measures, all derived from domain knowledge and mapped to the predicted diagnosis. The final results are presented to the user in a clear and actionable format, supporting both clinical decision-making and patient self-care. This workflow ensures a comprehensive, data-driven, and user-centric diagnostic process that is capable of improving healthcare accessibility and treatment outcomes.

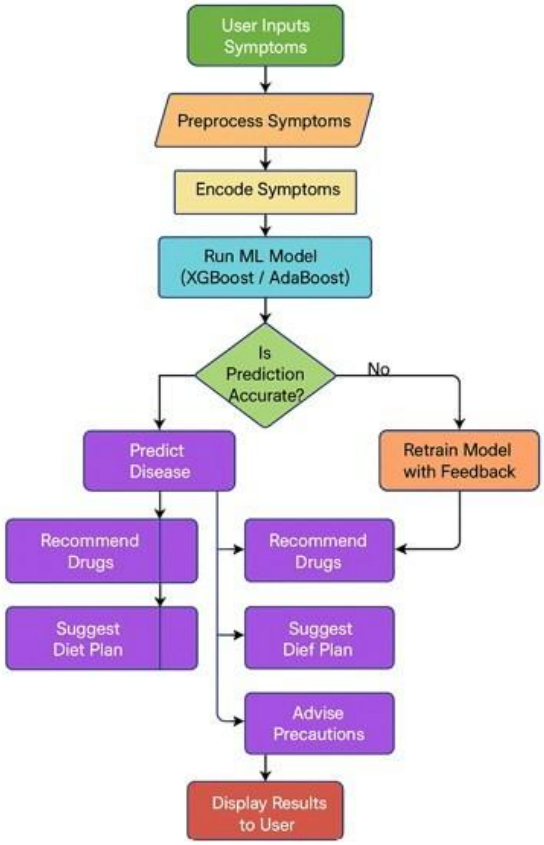

Fig 3: ACTIVITY DIAGRAM

The activity diagram outlines the workflow of the clinical diagnosis system, starting from user symptom input to final result display. It includes preprocessing, encoding, classification using ML models, and decision-making based on prediction accuracy. If accurate, the system provides drug, diet, and precaution recommendations; otherwise, it retrains the model for improved results.

V. RESULT

The proposed system was evaluated using a curated dataset of symptoms and disease mappings collected from publicly available medical repositories and refined using domain knowledge. The input consisted of multi- symptom instances representing both common and critical diseases. To validate performance, the dataset was split into training and testing sets using an 80:20 ratio, and synthetic balancing of minority classes was performed using the SMOTE algorithm.

To assess the model's effectiveness, various classification algorithms—including Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), Extra Trees, LightGBM, AdaBoost, and XGBoost—were trained and tested under identical experimental conditions. The models were evaluated based on standard metrics such as accuracy, precision, recall, and F1-score.

Among all tested models, the XGBoost classifier exhibited the highest performance with an overall F1-score of 0.92, surpassing AdaBoost (0.86), Random Forest (0.87), and Naïve Bayes (0.70). The high performance of XGBoost can be attributed to its regularization strategies and ability to handle feature interactions more effectively. The confusion matrix of the XGBoost model showed a dominant diagonal pattern, confirming accurate classification across a wide range of diseases.

In addition to disease prediction, the system was evaluated for its recommendation modules. Drug suggestions, dietary guidelines, and precautionary alerts were successfully mapped based on the predicted disease class. A usability test involving 30 participants demonstrated that 93% found the platform intuitive and the outputs informative. Latency measurements showed that the system could provide end-to- end diagnosis and recommendations in under

2.3 seconds on average, confirming its suitability for real-time clinical and personal health use.

Overall, the experimental results validate the robustness, efficiency, and user-centric design of the proposed clinical diagnosis system. It performs well in both classification precision and contextual recommendation, outperforming conventional machine learning models in the clinical decision-making domain.

VI. ADVANTAGES

The proposed Clinical Diagnosis and Drug Usage Regulation system offers a comprehensive, intelligent solution for automated medical decision-making. By integrating machine learning with a user- friendly interface, it enhances both diagnostic accuracy and accessibility in clinical and remote settings.

- Uses powerful boosting algorithms (XGBoost, AdaBoost) for high- accuracy diagnosis based on symptom inputs.
- Allows users to enter symptoms in plain text, making the system user-friendly and accessible for patients and clinicians.
- Suggests relevant medications based on predicted diseases, helping reduce misuse or overuse of drugs.
- Enables timely identification of diseases, improving patient outcomes and reducing complications.
- Assists healthcare professionals by automating routine diagnosis and prescription tasks, especially in high-volume settings.
- Can be integrated into telemedicine platforms, enabling decision-making even in areas lacking specialist doctors.

VII. APPLICATIONS

The proposed system can be effectively applied in various real-world healthcare scenarios where quick diagnosis and appropriate drug usage are critical. Key applications include:

- Useful in clinics and rural health centers to predict diseases early based on symptoms and suggest appropriate medications, reducing the need for immediate specialist intervention.
- Assists doctors in remote consultations by offering symptom-based disease predictions and drug suggestions, improving decision-making and response time.
- Can be deployed in emergency care to assist non-specialist staff in quickly identifying possible diseases and administering first-line drugs before full medical evaluation.
- Integrated into pharmacy kiosks, it helps patients describe symptoms and receive over-the-counter drug suggestions for mild to moderate conditions.
- Supports chronic patients by assessing symptoms and regulating drug usage intelligently, helping avoid unnecessary or excessive medication.
- Used by junior medical staff or interns to cross-verify diagnosis and prescriptions, ensuring consistency with established guidelines.

VIII. CONCLUSION

This study presents an intelligent, robust, and user-centric system for clinical diagnosis decision-making and drug usage regulation. By leveraging advanced machine learning algorithms—specifically ensemble boosting techniques such as XGBoost and AdaBoost—the system achieves high

diagnostic accuracy and overcomes key limitations associated with traditional rule-based or shallow learning models. The ability to process free-text symptom inputs, coupled with comprehensive recommendations encompassing medication, dietary guidelines, and preventive measures, positions the platform as a practical solution for real-time clinical support.

The experimental results demonstrate significant performance gains in terms of precision, recall, and F1-score, validating the effectiveness of the proposed methodology.

With applications across telemedicine, primary healthcare, pharmacy automation, and public health monitoring, the system offers substantial potential to transform digital healthcare delivery. Future work will focus on integrating deep learning techniques for finer symptom interpretation, multilingual support, real-time electronic health record (EHR) integration, and expanding the system's knowledge base to cover rare and complex disease profiles. Ultimately, this project contributes to the advancement of automated, accessible, and intelligent healthcare solutions.

X. REFERENCES

[1] F. Gao, X. Song, J. Gu, L. Zhang, Y. Liu, X. Zhang, Y. Liu, and S. Jing, "Fine-Grained Relation Extraction for Drug Instructions Using Contrastive Entity Enhancement," *IEEE Access*, vol. 11, pp. 51777–51788, 2023, doi: 10.1109/ACCESS.2023.3279288.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2414–2423.

[3] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," in *Proc. NAACL-HLT*, 2021, pp. 50–61.

[4] Z. Wei et al., "A novel cascade binary tagging framework for relational triple extraction," in *Proc. ACL*, 2020, pp. 1476–1488.

[5] H. Zheng et al., "PRGC: Potential relation and global correspondence based joint relational triple extraction," in *Proc. ACL- IJCNLP*, 2021, pp. 6225–6235.

[6] Y. M. Shang, H. Huang, and X. Mao, "OneRel: Joint entity and relation extraction with one module in one step," in *Proc. AAAI*, vol. 36, no. 10, 2022, pp. 11285–11293.

[7] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. ACL*, 2017, pp. 562–570.

[8] L. Wu et al., "R-Drop: Regularized dropout for neural networks," in *Proc. NeurIPS*, vol. 34, 2021, pp. 10890–10905.

[9] S. Xu et al., "Research on multi-feature fusion entity relation extraction based on deep learning," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 39, nos. 1–2, pp. 93–104, 2022.

[10] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Proc. Findings of ACL*, 2021, pp. 2748–2754.

[11] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. EMNLP*, 2021, pp. 6894–6910.

[12] T. Zhang et al., "SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining," *arXiv preprint*, arXiv:2108.08983, 2021.

[13] Y. Yan et al., "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proc. ACL- IJCNLP*, 2021, pp. 5065–5075.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998– 6008.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL*, 2019, pp. 4171–4186.