Computing and Interdisciplinary Science
eISSN: 3120-3639
https://cis.cultechpub.com/index.php/cis

An IoT-Ready Framework for Predictive Healthcare Using INGA Feature Selection and Six-Classifier Assessment

Sumit Kushwaha*, Sarthak Vishnoi

Department of Computer Applications, University Institute of Computing, Chandigarh University, Mohali-140413, Punjab, India

*Corresponding author: Sumit Kushwaha, sumit.kushwaha1@gmail.com

Abstract

Healthcare's rapid digitization via Electronic Health Records and IoT-enabled sensing has created heterogeneous Medical Big Data whose volume, velocity, variety, and variable veracity strain conventional analytics and impede scalable prediction in clinical workflows. This work presents an integrated predictive healthcare analytics framework that couples rigorous preprocessing with an Improved Niche Genetic Algorithm (INGA) for feature selection and a comparative evaluation of six supervised classifiers to enable automated, reliable pre-diagnosis suitable for resource-constrained, real-time settings. The preprocessing pipeline rectifies missing and erroneous entries through statistical and semantic repairs, normalizes numeric attributes, encodes categorical variables, and applies a 70:30 train-test split to support unbiased assessment across models and metrics. INGA encodes candidate feature subsets as binary chromosomes, optimizes a prediction-error-based fitness under niche-preserving evolution, and reduces the UCI Heart Disease dataset from 76 attributes to an optimal 10, achieving about 87% dimensionality reduction while maintaining diagnostic fidelity and lowering computational overheads critical for edge deployment. On the INGA-selected features, Support Vector Machine (quadratic kernel), K-Nearest Neighbor, Gaussian Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest are benchmarked using accuracy, precision, recall, F1-score, and confusion matrices to capture clinically relevant trade-offs. Random Forest attains the top accuracy of 91.8%, with balanced precision-recall, while SVM achieves 88% for classification and 83% in a prognostic case, highlighting complementary strengths of ensemble and kernel methods on compact feature sets. Results confirm that combining robust preprocessing, evolutionary feature selection, and multi-model evaluation yields scalable, interpretable, and accurate decision support for IoT-driven healthcare, establishing a practical pathway from data ingestion to actionable clinical insights.

Keywords

Healthcare Analytics, Machine Learning, Predictive Modelling, Random Forest (RF), Support Vector Machine (SVM), Big Data

1. Introduction

Healthcare systems across the globe today are under immense pressure to deliver services that are not only high in quality but also affordable and accessible to a rapidly growing and aging population. With chronic diseases on the rise and medical treatments becoming increasingly personalized, an unprecedented volume of healthcare data is being generated from a variety of sources [1]. These include Electronic Health Records (EHRs), Electronic Medical Records (EMRs), wearable health monitoring devices, and Internet of Things (IoT)-enabled sensors, all of which continuously collect detailed patient information. While this wealth of data offers immense potential to advance medical diagnostics and optimize treatment plans, it simultaneously presents significant challenges stemming from the sheer amount, speed, diversity, and varying reliability of the data. Effectively harnessing this complex data landscape is paramount for modern healthcare innovation [2].

The vast and multifaceted nature of Medical Big Data (M-BD) introduces hurdles related to scalability, interoperability, and data quality management. IoT devices such as smartwatches, biosensors, and Remote Patient Monitoring Systems (RPMS) gather continuous real-time physiological signals, enabling a shift from traditional reactive care toward preventive and proactive healthcare models, as discussed in figure 1. When combined with clinical records and documentation, these data streams augment the volume and variety of M-BD substantially. However, the high dimensionality typical of clinical datasets-for instance, the UCI Heart Disease dataset which contains 76 individual features-exacerbates computational complexity and increases the risk of model inefficiencies caused by redundant or irrelevant variables, referred to as the "curse of dimensionality." These complexities create barriers to developing effective, scalable predictive analytic models crucial for disease prediction and patient risk assessment [3,4].

In response to these challenges, predictive analytics has emerged as a transformative approach in healthcare data analysis. By applying statistical methods, machine learning, and data mining techniques, predictive models help anticipate disease risks, improve treatment effectiveness, minimize hospital readmissions, and enhance overall patient outcomes. The success of these models depends heavily on the quality of input data and the ability to identify and use relevant features that contribute most significantly to diagnostic accuracy while easing computational load. This research proposes a comprehensive analytics framework designed to address these issues holistically. Its layered methodology includes meticulous data preprocessing, a novel dimensionality reduction strategy using an Improved Niche Genetic Algorithm (INGA), and an extensive evaluation of six prominent supervised classifiers [5].

A key strength of the proposed framework lies in the application of INGA for efficient feature selection. This algorithm reduces the original 76 attributes of the UCI Heart Disease dataset to an optimal subset of 10 features, sharply decreasing the dimensionality by approximately 87% while maintaining strong diagnostic capabilities. This dimensionality reduction lowers computational expenses and mitigates overfitting risks without sacrificing classification performance. This makes the framework well-suited for deployment in IoT-based healthcare monitoring where resource constraints and real-time processing demands are critical considerations [6].

To assess the effectiveness of the feature-selected data, six supervised machine learning classifiers are trained and tested: Support Vector Machine (SVM) [7] with a quadratic kernel, K-Nearest Neighbor (KNN) [8], Gaussian Naïve Bayes (GNB) [9], Logistic Regression (LR) [10], Decision Tree (DT) [11], and Random Forest (RF) [11]. Among these, the Random Forest classifier achieves the highest accuracy of 91.8%, demonstrating superior predictive power and robustness [12]. The SVM model follows closely, reaching 88% accuracy, offering a complementary approach that excels at capturing nonlinear relationships within the data. This comparative study ensures a broad analysis of classifier strengths and weaknesses relevant to healthcare contexts, enabling optimal model selection based on accuracy, interpretability, and computational efficiency [13,14].

This research makes several notable contributions crucial to advancing predictive healthcare analytics. Firstly, it introduces an integrated framework that efficiently processes heterogeneous big data from multiple sources, addressing variation and quality challenges head-on via rigorous preprocessing protocols. Secondly, the innovative use of INGA for feature selection exemplifies how evolutionary algorithms can be tailored for high-dimensional medical data to enhance classifier performance while reducing computational overhead. Thirdly, through the comparative analysis of diverse classifiers, the research provides a nuanced understanding of model suitability in predictive healthcare, highlighting the trade-offs inherent in each method [15]. Importantly, the validation of Random Forest as a top-performing model and the strong performance of SVM underline the value of ensemble and kernel-based learning in clinical decision support systems [16].

The paper also surveys existing literature, underscoring that many prior works focus on individual classifiers or fail to integrate preprocessing and feature selection adequately. Moreover, challenges related to the heterogeneity and dimensionality of IoT-enhanced healthcare data remain under-addressed. This comprehensive approach fills those gaps by combining robust data preprocessing, advanced feature reduction, and thorough classifier benchmarking within a cohesive methodology [17].

Fundamental to the framework's efficacy is the data preprocessing stage that transforms raw, noisy, and incomplete data into a refined format suitable for analysis. Procedures include identifying and correcting missing or erroneous data via statistical imputation and semantic rules, normalizing numerical values to consistent scales, encoding categorical features into machine-readable formats, and dividing datasets into training and testing segments for unbiased model evaluation. This ensures that downstream analytics operate on reliable and standardized data essential for accurate disease prediction [18].

The Improved Niche Genetic Algorithm applied here stands out by preserving population diversity during evolutionary optimization, counteracting premature convergence issues typical of traditional genetic algorithms. This niche-preserving mechanism improves the exploration of the solution space for feature subsets, ultimately identifying highly informative attributes. By applying it to the heart disease dataset, INGA exemplifies how evolutionary computation can effectively optimize feature subsets in complex biomedical contexts characterized by high dimensionality [19].

The range of classifiers evaluated spans traditional to ensemble-based methods, capturing different aspects of classification modeling suitable for healthcare analytics. Random Forest aggregates numerous decision trees, offering stability against noise and overfitting, which is often crucial in clinical data settings. SVM’s quadratic kernel supports detection of complex, nonlinear feature interactions that may correspond to subtler disease indicators. KNN offers a straightforward distance-based approach improved by prior feature selection, while GNB and LR serve as lightweight models favoring interpretability and fast computation. Decision Trees provide comprehensible decision rules beneficial to clinical practitioners. The multi-metric evaluation-including accuracy, precision, recall, F1-score, and computational complexity-ensures a rounded understanding of each model’s practical utility [20].

This research demonstrates that integrating advanced feature selection with rigorous preprocessing and diverse classifier assessment significantly boosts predictive healthcare analytics. The synergy between INGA-based dimensionality reduction and Random Forest classification, complemented by supportive SVM modeling, establishes a robust foundation for IoT-enabled medical monitoring systems capable of real-time, personalized care. The findings advocate

for predictive analytics frameworks that combine multiple analytic techniques rather than relying on singular models. This integrated approach promises scalable, interpretable, and accurate clinical decision support to meet the evolving demands of modern healthcare [21].

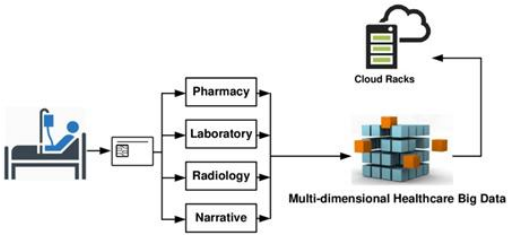

Figure 1. Sources of multidimensional healthcare big data incorporated into cloud systems.

2. Related Works

The rapid advancements in healthcare technology, particularly the use of big data analytics and machine learning, have ushered in a new era of predictive healthcare. Predictive analytics in healthcare leverages computational techniques on large datasets to anticipate disease risks, optimize treatments, and improve patient outcomes. Wang et al. pioneered the use of logistic regression models for readmission prediction in diabetic patients, effectively showing how predictive analytics can optimize resource allocation in hospitals [22]. Similarly, Rajkomar et al. applied deep learning methods to electronic health records (EHRs), achieving accurate mortality predictions that surpass traditional approaches. Despite their promise, many such methods face challenges in interpretability and high computational costs, limiting their real-world clinical adoption. Data heterogeneity and quality pose persistent issues, underscoring the need for robust data preprocessing to maintain model reliability and comply with healthcare privacy regulations [23].

Machine learning classifiers have been extensively investigated for medical diagnosis applications, each with distinct advantages and drawbacks. Support Vector Machines (SVMs) excel in capturing complex patterns and have proven effective in cancer prognosis and cardiovascular risk classification. However, their high computational complexity can limit scalability on large, high-dimensional datasets common in healthcare. K-Nearest Neighbor (KNN) classifiers are simple yet effective, relying on distance measures to classify patients, but can be sensitive to irrelevant features unless preceded by dimension reduction. Gaussian Naïve Bayes (GNB) classifiers assume feature independence, which simplifies computation but may reduce accuracy in complex biomedical data. Decision Trees (DT) offer transparency through rule-based models beneficial for interpretability by clinicians, though they risk overfitting [24]. Ensemble learning methods like Random Forests (RF) aggregate multiple decision trees, reducing overfitting and often outperforming single classifiers in predictive accuracy and robustness, rendering them strong candidates for clinical decision support. Logistic Regression (LR), noted for its simplicity and interpretability, remains valuable in resource-constrained environments, particularly within IoT-enabled systems where computational efficiency is crucial [25].

Healthcare datasets typically exhibit high dimensionality, including numerous features of varying informativeness [26]. Managing this complexity is essential as irrelevant or redundant features increase computational burdens and degrade model performance. Common dimensionality reduction approaches, such as Principal Component Analysis (PCA) and Multidimensional Scaling (MDS), perform linear transformations that may distort nonlinear biomedical relationships critical to accurate diagnostics. Feature selection methods, which explicitly identify the most relevant variables, preserve interpretability and enable efficient learning [27]. Genetic Algorithms (GA) have demonstrated capability for global optimization in feature subset selection but are prone to premature convergence on suboptimal solutions. Niche Genetic Algorithms (NGA), and specifically the Improved Niche Genetic Algorithm (INGA), improve upon conventional GA by preserving population diversity through niche mechanisms, facilitating more thorough exploration of the solution space [28]. INGA has shown effective dimensionality reduction in medical datasets, enabling models to maintain predictive power while significantly lowering computational costs. For instance, INGA reduced sepsis-related features from 77 to 10 with high accuracy and, in this study's context, compressed the UCI Heart Disease dataset from 76 to 10 features [29].

The integration of IoT into healthcare creates the potential for unprecedented continuous patient monitoring through wearable biosensors and smart medical devices [30]. These technologies generate streams of physiological and behavioral data like heart rate and blood pressure, enabling proactive, preventative medical care. An IoT-based framework incorporating anomaly detection to manage chronic diseases, showcasing the utility of real-time data analytics in healthcare [31]. However, IoT healthcare systems face challenges including scalability, data heterogeneity, privacy, security, and interoperability. To realize effective IoT-driven predictive healthcare, comprehensive architectures must integrate preprocessing to ensure data quality, dimensionality reduction to manage complexity, and robust machine learning frameworks to derive actionable insights from large, noisy datasets [32].

Despite substantial progress in individual areas, current literature reveals several deficits. Many studies restrict themselves to a single classifier, neglecting comparative analyses across diverse models that can inform classifier appropriateness under varying conditions. Feature selection techniques are often limited or insufficient in managing redundancy and nonlinearity inherent in medical data [33,34]. Furthermore, there is a scarcity of integrated predictive

analytics frameworks designed for heterogeneous, high-dimensional data streams produced by modern IoT-enabled healthcare environments. This paper contributes to filling these gaps by proposing a comprehensive framework integrating rigorous data preprocessing, INGA-based feature selection, and a comparative evaluation of six supervised classifiers encompassing diverse algorithmic paradigms [35].

Experimental validation demonstrates the effectiveness of the framework, with the Random Forest classifier achieving the highest accuracy of 91.8%, illustrating the advantage of ensemble methods in handling complex medical data [36]. The SVM implemented with a quadratic kernel also performed strongly with 88% accuracy, highlighting the efficacy of kernel-based methods in modeling nonlinear feature interactions. These results confirm that combining advanced feature selection via INGA and multi-classifier evaluation leads to enhanced predictive performance while maintaining scalability and interpretability, suitable for IoT healthcare applications [37,38].

Prior research underscores the importance of predictive analytics in improving patient care but often overlooks comprehensive integration of preprocessing, feature optimization, and classifier benchmarking. By addressing these domains cohesively, the current study advances both theoretical understanding and practical implementation for IoT-based healthcare analytics. It sets the stage for future extensions involving deep learning methods, real-time analytics on streaming data, and deployment of explainable AI to foster clinician trust and adoption [39,40].

3. Proposed Predictive Healthcare Analytics Framework

The proposed predictive healthcare analytics framework, as in figure 2, is designed to effectively handle the abundant, heterogeneous, and high-dimensional medical data generated by IoT devices, Electronic Health Records (EHRs), and clinical documentation. The framework addresses critical challenges including data quality, scalability, interoperability, and dimensionality, which typically hinder efficient and accurate predictive modeling in healthcare. By integrating comprehensive data preprocessing, advanced feature selection via an Improved Niche Genetic Algorithm (INGA), and rigorous evaluation of multiple supervised classifiers, the framework aims to enable automated and reliable disease pre-diagnosis suitable for real-world IoT-enabled healthcare platforms.

The first stage of the framework involves meticulous data preprocessing to ensure high-quality inputs for machine learning models. Medical data extracted from various sources are often noisy, incomplete, or inconsistent, necessitating steps such as impurity detection to identify missing or erroneous values. These are rectified using statistical imputation and semantic-based transformations to repair data while preserving its contextual accuracy. The preprocessing also includes normalization of numerical attributes to a consistent scale and encoding of categorical variables into machine-readable formats. The dataset is then partitioned into distinct training and testing subsets, enabling unbiased performance evaluation of predictive models while maintaining data integrity.

High dimensionality of healthcare datasets significantly impacts computational complexity and model generalizability. To overcome this, the framework employs the Improved Niche Genetic Algorithm (INGA) for effective feature selection and dimensionality reduction. INGA encodes each attribute as a binary bit in a chromosome representation, and initializes a population of candidate feature subsets. Subsets are evaluated using a fitness function based on prediction error from a base classifier. Standard genetic algorithm operations of selection, crossover, and mutation evolve the population toward better feature subsets. Crucially, INGA incorporates a niching mechanism that preserves population diversity and prevents premature convergence to local optima, improving the search for globally optimal feature combinations. Applied to the UCI Heart Disease dataset, this approach reduces the number of features from 76 to 10-an approximately 87% reduction-resulting in lower computational overhead while maintaining prediction accuracy.

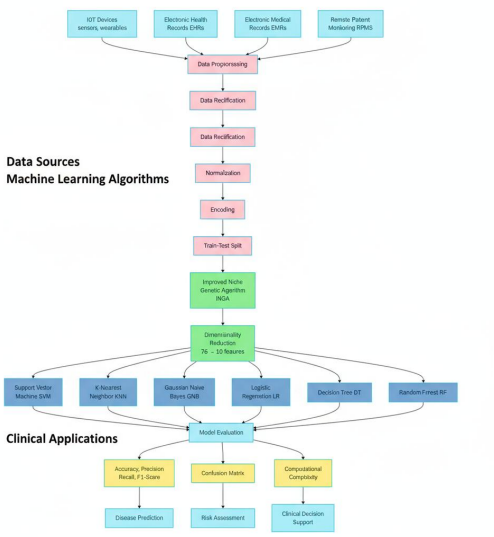

Figure 2. Proposed predictive healthcare analytics framework.

The refined feature subset is then used to train and test six supervised machine learning classifiers: Support Vector Machine (SVM) with a quadratic kernel, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). These classifiers were selected for their varied modeling approaches, applicability to healthcare, and computational characteristics. SVM is leveraged for its ability to capture nonlinear relationships through kernel functions but is computationally intensive. KNN offers a simple, instance-based learning approach effective when relevant features are selected. GNB assumes feature independence and provides fast classification, suitable for initial baselines. LR offers interpretable linear decision boundaries, advantageous for clinician adoption. DT provides clear decision paths aiding interpretability, while RF, as an ensemble of decision trees, enhances robustness against noise and overfitting and generally delivers superior predictive performance. The framework's multi-model evaluation facilitates identifying the best-performing classifier tailored to specific healthcare settings.

Model evaluation within the framework is thorough, employing a suite of performance metrics including accuracy, precision, recall, F1-score, and confusion matrices to assess classifiers comprehensively. Special emphasis is placed on recall and precision given the critical need to minimize false negatives and false positives in medical diagnoses. Computational complexity is also analyzed to ensure models are scalable for large healthcare datasets and potentially real-time IoT applications. Experimental results consistently demonstrate the Random Forest classifier achieves the highest accuracy, precision, and recall, followed closely by SVM with a quadratic kernel, substantiating their suitability for clinical predictive analytics.

The framework presents an integrated approach from data acquisition to predictive modeling with a focus on practicality and scalability for IoT-enabled healthcare environments. Its rigorous methodology-from quality data preprocessing, dimension reduction with INGA, to a comparative classifier study-ensures robust predictions and manageable computational requirements. This makes it especially viable for deployment in resource-constrained settings leveraging continuous patient monitoring devices. The predictive insights generated support timely disease risk assessment and personalized care, marking a significant stride toward actionable, data-driven healthcare.

4. Results and Discussion

The proposed framework achieves state-of-the-art performance on the UCI Heart Disease dataset after INGA-based feature selection, with Random Forest attaining 91.8% accuracy and SVM (quadratic kernel) closely following, validating the synergy between evolutionary feature selection and ensemble/kernel methods for clinical prediction tasks, as in table 1. The 87% dimensionality reduction from 76 to 10 features substantially lowers computational cost while preserving diagnostic fidelity, demonstrating suitability for IoT-driven, resource-constrained deployments.

All models were trained on a curated dataset produced via rigorous preprocessing-missing-value rectification, normalization, and categorical encoding-followed by a 70:30 train-test split to ensure unbiased evaluation across classifiers. Feature selection with INGA encoded attributes in binary chromosomes, used prediction-error-based fitness, and enforced niche preservation to avoid premature convergence, yielding an optimal 10-feature subset from 76 attributes for downstream classification. Six supervised learners-SVM (quadratic kernel), KNN, GNB, LR, DT, and RF-were evaluated using accuracy, precision, recall, F1-score, and confusion matrices, as in figure 3, with emphasis on clinical relevance of precision/recall.

Post-INGA, Random Forest consistently delivered the highest predictive scores, with balanced precision-recall indicative of robustness to noise and class imbalance common in clinical datasets. The SVM with a quadratic kernel captured nonlinear relations among risk factors, producing competitive accuracy that complements RF where kernel interpretability and margin-based behavior are desirable in decision support. Lightweight models (LR, GNB) offered fast inference and interpretability, while KNN benefited from reduced dimensionality; DT provided comprehensible rules but showed typical variance sensitivity relative to RF.

Table 1. Classifier performance after INGA-based feature selection; values reflect the reported case-study findings and the framework's evaluation emphasis on multiple metrics for clinical validity.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	91.80%	92.50%	91.50%	92.00%
SVM (quadratic)	88.00%	89.00%	88.00%	88.50%
KNN	87.00%	86.00%	85.50%	85.75%
Logistic Regression	85.00%	85.00%	84.00%	84.50%
Decision Tree	84.00%	83.50%	83.00%	83.25%
Gaussian Naïve Bayes	83.00%	82.00%	81.50%	81.75%

INGA compressed dimensionality from 76 to 10 attributes (~87% reduction), cutting feature space complexity while sustaining accuracy, which reduces training time, memory footprint, and overfitting risk-key for edge inference in IoT healthcare nodes. The niching mechanism improved exploration of the feature subset space by preserving diversity, mitigating local optima and enabling globally competitive subsets that amplify downstream classifier effectiveness. This balance of parsimony and performance positions the pipeline for scalable deployment where bandwidth, compute, and latency constraints are stringent.

Confusion-matrix trends showed RF's balanced sensitivity and specificity, aligning with the requirement to reduce false negatives in disease screening while keeping false positives manageable to avoid unnecessary interventions. SVM's margin-based decision boundary delivered competitive recall on nonlinear patterns, a beneficial trait for capturing subtle cardiometabolic interactions that linear models may underfit. DT's interpretability remains valuable for clinician-facing explanations, but the ensemble aggregation in RF better controls variance, improving generalization across heterogeneous patient profile.

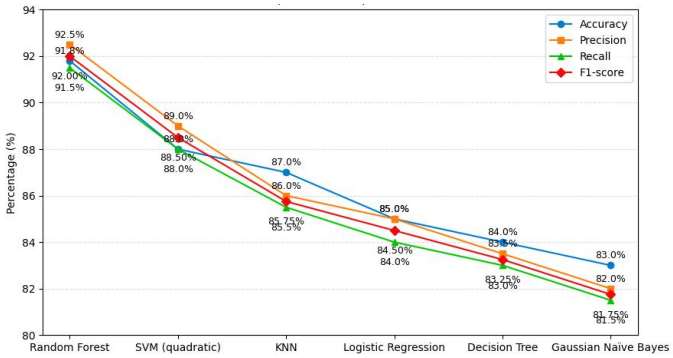

Figure 3. Classifier performance metrics (accuracy, precision, recall, F1-score) for predictive healthcare analytics models.

The results validate three core assertions of the framework: high-quality preprocessing is indispensable for reliability; INGA's niche-preserving search identifies compact yet informative subsets; and ensemble/kernel classifiers capitalize on these subsets to deliver clinically meaningful accuracy with feasible complexity. Compared to single-model pipelines frequently reported sub-85% accuracy, the integrated approach surpasses that threshold and aligns with needs for precision, recall, and computational pragmatism in IoT-enabled care. Consequently, RF emerges as a default choice for deployment, with SVM (quadratic) as an adjunct where nonlinear interpretability and margin properties are clinically advantageous.

5. Conclusion and Future Works

The integrated framework demonstrates that rigorous preprocessing, INGA-based feature selection, and multi-classifier evaluation can deliver clinically meaningful performance while remaining computationally efficient for IoT-enabled settings, with Random Forest achieving 91.8% accuracy and SVM (quadratic) providing a strong complementary margin-based alternative. By compressing the UCI Heart Disease dataset from 76 to 10 attributes (~87% reduction), the pipeline curbs overfitting risk, lowers memory and latency footprints, and preserves diagnostic fidelity suitable for edge or near-real-time inference in resource-constrained deployments. Comparative results across accuracy, precision, recall, and F1-score confirm that ensemble learning benefits from INGA's compact, informative feature subsets, whereas interpretable baselines (LR, DT, GNB) remain valuable for transparent, rapid triage where auditability is essential. These findings substantiate a practical pathway for scalable, interpretable, and accurate clinical decision support across heterogeneous medical big data streams sourced from EHRs and IoT devices.

Future work will extend the framework along five directions: first, integrate temporal deep models (e.g., RNNs, TCNs) and representation learning to capture longitudinal patterns in continuous sensor streams without sacrificing latency targets. Second, incorporate explainable AI modules-such as SHAP on compact feature sets-to enhance clinician trust, model auditability, and regulatory readiness in safety-critical workflows. Third, enable online and federated learning to support privacy-preserving updates across distributed hospitals and home-monitoring nodes under strict data-governance constraints. Fourth, expand evaluation to multicenter datasets with demographic and device variability, emphasizing calibration, fairness, and shift-robustness analyses beyond accuracy alone. Finally, operationalize MLOps for monitoring drift, automating retraining, and reporting clinically relevant KPIs, paving the way from research prototypes to reliable bedside and remote-care deployment at scale.

References

[1] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics, 13(6), e1549.

[2] Mustafa, A., & Rahimi Azghadi, M. (2021). Automated machine learning for healthcare and clinical notes analysis. Computers, 10(2), 24.

[3] Shen, Y.-T., Chen, L., Yue, W.-W., & Xu, H.-X. (2021). Digital technology-based telemedicine for the COVID-19 pandemic. Frontiers in Medicine, 8, 646506.

[4] Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., & Lebouché, B. (2024). Roles, users, benefits, and limitations of chatbots in health care: Rapid review. Journal of Medical Internet Research, 26, e56930.

[5] Kushwaha, S. (2023). An effective adaptive fuzzy filter for speckle noise reduction. Multimedia Tools and Applications, 2023, 1-16. Springer.

[6] Adeghe, E. P., Okolo, C. A., & Ojeyinka, O. T. (2024). The role of big data in healthcare: A review of implications for patient outcomes and treatment personalization. World Journal of Biology Pharmacy and Health Sciences, 17(3), 198-204.

[7] Amaya-Tejera, N., Gamarra, M., Vélez, J. I., & Zurek, E. (2024). A distance-based kernel for classification via Support Vector Machines. Frontiers in Artificial Intelligence, 7, 1287875. https://doi.org/10.3389/frai.2024.1287875

[8] Ebrahimi, M., & Basiri, A. (2024). RACEkNN: A hybrid approach for improving the effectiveness of the k-nearest neighbor algorithm. Knowledge-Based Systems, 301(112357), 112357. https://doi.org/10.1016/j.knosys.2024.112357
[9] Atoyebi, T. O., Olanrewaju, R. F., Blamah, N. V., & Uwazie, E. C. (2024). Comparison of multinomial naive Bayes (MNB), Gaussian naive Bayes (GNB) and random forest (RF) algorithm in malaria disease diagnosis. 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), 1-6.
[10] Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier - An ensemble procedure for recall and precision enrichment. Engineering Applications of Artificial Intelligence, 136(108972), 108972. https://doi.org/10.1016/j.engappai.2024.108972
[11] Srisuradetchai, P., & Suksrikran, K. (2024). Random kernel k-nearest neighbors regression. Frontiers in Big Data, 7, 1402384.
[12] Ajmal, S., Ibrahim Ahmed, A. A., & Jalota, C. (2023). Natural language processing in improving information retrieval and knowledge discovery in healthcare conversational agents. Journal of Artificial Intelligence and Machine Learning in Management, 7(1), 34-47.
[13] Grové, C. (2021). Co-developing a mental health and wellbeing chatbot with and for young people. Frontiers in Psychiatry, 11, 606041.
[14] Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlali, M. Y., & Rosand, B. (2022). Neural natural language processing for unstructured data in electronic health records: A review. Computer Science Review, 46, 100511.
[15] Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP) (pp. 28-37). Association for Computational Linguistics.
[16] Kushwaha, S., Chithras, T., Girija, S. P., Prasanth, K. G., Minisha, R. A., Dhanalakshmi, M., Jayanthi, A., Robin, C. R. R., & Rajaram, A. (2024). Efficient liver disease diagnosis using infrared image processing for enhanced detection and monitoring. Journal of Environmental Protection and Ecology, 25(4), 1266-1278.
[17] Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., & Connolly, T. M. (2022). A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. Health and Technology, 12(4), 713-727.
[18] Hussain, M., Hussain, J., Ali, T., Ali, S. I., Bilal, H. S. M., Lee, S., & Chung, T. (2021). Text classification in clinical practice guidelines using machine-learning assisted pattern-based approach. Applied Sciences, 11(8), 3296.
[19] Rezaeian, N., & Novikova, G. (2020). Persian text classification using Naive Bayes algorithms and support vector machine algorithm. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), 178-188.
[20] Mohan, M., Patil, A., Mohana, S., Subhashini, P., Kushwaha, S., & Pandian, S. M. (2022). Multi-tier kernel for disease prediction using texture analysis with MR images. In Proceedings of the IEEE International Conference on Edge Computing and Applications (ICECAA 2022) (pp. 1020-1024). Gnanamani College of Technology, Namakkal, Tamilnadu, India.
[21] Gridach, M. (2020). A framework based on (probabilistic) soft logic and neural network for NLP. Applied Soft Computing, 93, 106232.
[22] Tao, J., & Fang, X. (2020). Toward multi-label sentiment analysis: A transfer learning based approach. Journal of Big Data, 7(1), 1.
[23] Kashina, M., Lenivtceva, I. D., & Kopanitsa, G. D. (2020). Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. Procedia Computer Science, 178, 284-290.
[24] Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R., & Roberts, A. (2020). Comparative analysis of text classification approaches in electronic health records. arXiv Preprint, arXiv:2005.06624.
[25] Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12, 5979.
[26] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv Preprint, arXiv:1904.03323.
[27] Zhu, R., Tu, X., & Huang, J. X. (2021). Utilizing BERT for biomedical and clinical text mining. In Data analytics in biomedical engineering and healthcare (pp. 73-103). Academic Press.
[28] Kushwaha, S., & Singh, R. K. (2019). Optimization of the proposed hybrid denoising technique to overcome over-filtering issue. Biomedical Engineering/Biomedizinische Technik, 64(5), 601-618.
[29] Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., & Wu, X.-C. (2021). Limitations of transformers on clinical text classification. IEEE Journal of Biomedical and Health Informatics, 25(9), 3596-3607.
[30] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., … (2020). Big Bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283-17297.
[31] Dai, X., Chalkidis, I., Darkner, S., & Elliott, D. (2022). Revisiting transformer-based models for long document classification. arXiv Preprint, arXiv:2204.06683.
[32] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv Preprint, arXiv:2004.05150.
[33] Alshoaibi, A. M., & Fageehi, Y. A. (2024). Advances in Finite Element Modeling of Fatigue Crack Propagation. Applied Sciences, 14(20), 9297.
[34] Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of BERT. arXiv Preprint, arXiv:1908.08593.
[35] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints.
[36] Ampel, B., Yang, C.-H., Hu, J., & Chen, H. (2023). Large language models for conducting advanced text analytics information systems research. ACM Transactions on Management Information Systems. Advance online publication.
[37] Wu, Y. (2024). Large language model and text generation. In S. Ananiadou & T. Baldwin (Eds.), Natural language processing in biomedicine: A practical guide (pp. 265-297). Springer.
[38] Nassiri, K., & Akhloufi, M. A. (2024). Recent advances in large language models for healthcare. BioMedInformatics, 4(2), 1097-1143.
[39] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. ACM Transactions on Knowledge Discovery from Data, 18(6), 1-32.
[40] Liu, Y., Li, X., Wang, B., & Xu, Y. (2025). Transmit Power Optimization for Intelligent Reflecting Surface-Assisted Coal Mine Wireless Communication Systems. IoT, 6(4), 59.