



DISSEMINATING THE RISK FACTORS WITH ENHANCEMENT IN PRECISION MEDICINE USING COMPARATIVE MACHINE LEARNING MODELS FOR HEALTHCARE DATA

¹ Dr. Yasaswini Vanapalli, ² K.Madhupranavi sai, ³ K.V.S.M. Bhargavi, ⁴ S.N.R.L.D.Swathi

¹ Associate Professor, Department of CSE (Cyber Security), School of CSE, Malla Reddy Engineering college for women, Hyderabad, India.

^{2,3,4} Students, Department of Computer Science & Engineering (IOT), School of CSE, Malla Reddy Engineering college for women, Hyderabad, India.

ABSTRACT

The field of healthcare analytics, as an expanding discipline that integrates data analysis, machine learning, and clinical expertise, is seen to hold great promise for improving patient outcomes and the overall delivery of healthcare services. With the increasing availability of Electronic Health Records (EHRs), a wealth of healthcare data has emerged, presenting opportunities to enhance disease prediction and personalize treatments. The objective of the research is to develop and evaluate machine-learning models for predicting cancer, diabetic, diabetic retina, and heart-related outcomes using demographic and clinical data from Electronic Health Records (EHRs). Through thorough testing on diverse datasets, the study aims to assess the performance of these models in terms of accuracy, precision, and recall metrics, with the ultimate goal of advancing disease prediction and enhancing patient outcomes within the field of healthcare analytics. The proposed model demonstrates high accuracy, particularly in predicting cancer (97.080%) and diabetic (97.33%) outcomes using Support Vector Machines (SVM) and Decision Trees. Additionally, logistic regression achieves a notable accuracy of 76.521% for diabetic retina dataset, while Decision Trees exhibit 86.419% accuracy for heart-related predictions. SVM accuracy for Pima diabetic dataset stands at 79.746%. To assess the model's performance, thorough testing was conducted on a diverse and extensive dataset, employing a combination of accuracy, precision, and recall metrics. This research represents a substantial contribution to the field of healthcare analytics, emphasizing the potential of machine learning to advance disease prediction and, ultimately, enhance patient outcomes.

Keywords: Precision Medicine, Healthcare Data Analytics, Risk Factor Analysis, Comparative Machine Learning, Predictive Modeling, Deep Learning, Supervised Learning, Unsupervised Learning, Feature Selection, Model Validation, Explainable AI (XAI), Clinical Decision Support Systems, Electronic Health Records (EHR), Biomedical Informatics, Disease Risk Prediction, Data-Driven Healthcare, Computational Medicine, Patient-Specific Prediction, Digital Health, AI in Healthcare.

1.INTRODUCTION

Precision medicine represents a transformative approach to healthcare that focuses on tailoring medical treatment to the individual characteristics of each patient. It leverages large-scale healthcare data, including clinical, genomic, and lifestyle information, to develop personalized diagnostic and therapeutic strategies. However, the explosion of heterogeneous healthcare data has made it difficult to extract meaningful insights and disseminate the underlying risk factors that

contribute to disease progression. Traditional analytical methods often fall short due to their inability to handle high-dimensional, non-linear, and correlated data. This limitation has led to the adoption of machine learning (ML) and deep learning (DL) models, which can uncover hidden patterns and correlations in complex datasets.

The primary objective of this research is to enhance the dissemination and identification of risk factors in precision medicine by employing comparative machine learning models on



healthcare data. By comparing various ML algorithms such as Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Machines (GBM), and Deep Neural Networks (DNN), this study aims to determine the optimal model for disease prediction and patient stratification. Furthermore, explainable AI techniques are integrated to improve the interpretability of the model outcomes, ensuring that clinicians can trust and understand the predictive results.

II. LITERATURE SURVEY

Over the past decade, numerous studies have demonstrated the potential of artificial intelligence in medical data analysis. Johnson et al. (2021) applied ensemble learning methods to cardiovascular datasets, showing that hybrid algorithms outperform traditional statistical models in risk prediction. Li et al. (2022) employed deep learning frameworks for oncology, proving that convolutional and recurrent neural networks can detect cancer markers with high sensitivity and specificity. Similarly, Kumar and Singh (2023) used SVM and Random Forest for diabetes detection, emphasizing the role of preprocessing and feature engineering in improving model reliability. Rahman et al. (2024) focused on integrating Explainable AI (XAI) into healthcare systems, highlighting its significance in bridging the gap between data-driven predictions and clinical reasoning. While these studies achieved substantial advancements, most of them concentrated on a single model or disease domain. Few studies have compared multiple ML models across heterogeneous healthcare datasets to identify and disseminate key risk factors applicable to precision medicine. Moreover, existing approaches often neglect the need for explainability, which is crucial for real-world clinical adoption. Therefore, a comparative, transparent, and scalable system is necessary to advance data-driven healthcare research.

III. EXISTING SYSTEM

Existing systems for healthcare prediction primarily rely on isolated machine learning models, often limited to specific diseases or static datasets. These systems face several challenges, including incomplete data preprocessing, class imbalance, and the lack of feature interpretability. Many current implementations do not handle missing data effectively, leading to biased predictions. Furthermore, models such as logistic regression or single decision trees fail to capture complex interactions between genetic, behavioral, and environmental variables.

Another major limitation is the absence of comparative evaluation—most systems do not benchmark multiple models under identical conditions, making it difficult to identify the best-performing algorithms. Additionally, existing solutions rarely employ explainable AI techniques, which results in “black box” predictions that clinicians cannot interpret. Finally, the lack of integration with Electronic Health Records (EHRs) restricts data sharing and interoperability across healthcare institutions, limiting scalability and real-time use in clinical environments.

IV. PROPOSED SYSTEM

The proposed system introduces an end-to-end, comparative machine learning framework designed to enhance risk factor identification and prediction accuracy in precision medicine. The system begins by collecting and aggregating healthcare data from multiple sources such as EHRs, wearable devices, and biomedical datasets. The data preprocessing module standardizes the data by handling missing values, removing noise, and normalizing attributes. This ensures that models receive clean and consistent input for training.

Next, feature selection techniques—such as correlation analysis, chi-square testing, and mutual information—are applied to identify the most influential health indicators.



These features are then used to train multiple ML algorithms including Support Vector Machines, Random Forests, Gradient Boosting Machines, and Deep Neural Networks. Each model undergoes rigorous evaluation using performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC. The comparative analysis identifies the model that best balances prediction accuracy and computational efficiency. To enhance transparency, the system integrates Explainable AI (XAI) methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These tools explain the contribution of each feature to the final prediction, allowing clinicians to understand why a particular patient is classified as high or low risk. The system's modular design supports cloud deployment, ensuring scalability and accessibility for multi-institutional healthcare system

V.SYSTEM ARCHITECTURE

The proposed architecture is an end-to-end pipeline that ingests heterogeneous healthcare data and turns it into clinician-ready risk insights. Structured EHR records, wearable/IoT telemetry, omics repositories (genomics/proteomics), and public-health/SDoH sources first pass through a secure ingestion and governance layer that enforces access control, PHI masking, schema validation, and audit logging. A preprocessing and harmonization stage then performs de-identification, imputation, normalization, outlier handling, and categorical encoding to produce analysis-ready tables. Features are engineered and persisted in a versioned feature store (statistical aggregates, signal windows, and learned embeddings), backed by a data lake for lineage and reproducibility. A modeling hub trains a comparative suite of algorithms—SVM, Random Forest, Gradient Boosting, and DNN—under consistent cross-validation with stratified

splits; the evaluation/selection unit ranks models by AUC, F1, precision-recall, calibration, and cost-sensitive metrics. An XAI layer (global and local SHAP/LIME) attributes predictions to clinical variables and generates patient-specific narratives, which feed a risk-factor dissemination service producing ranked factor lists, cohort summaries, and exportable reports. The best model is deployed behind a serving API (batch and real-time) with a clinician dashboard for risk scores, trends, what-if analysis, and alerting. MLOps & monitoring track data drift, retraining triggers when thresholds are breached. Finally, hybrid cloud–edge deployment allows low-latency scoring at hospital edge nodes while retaining heavy training and governance in the cloud, optimizing cost, privacy, and resilience.



Fig 5.1 System Architecture

VI.IMPLEMENTATION



Fig 6.1 Prediction of Animal Detection



Comparative Machine Learning Model Analysis

A comparative analysis was performed across four distinct machine learning algorithms. The primary metrics of interest for this high-stakes risk prediction task are "Area Under the ROC Curve (AUC)" and "F1-Score".

Model Name	Accuracy (Avg)	AUC (Avg)	F1 Score (Avg)	Sensitivity (Recall)
"Random Classifier"	0.555	~0.500*	~0.500*	0.555
Random Forest	0.875	0.910	0.870	0.880
Support Vector Machine (SVM)	0.700	0.880*	0.820*	0.900
Logistic Regression	0.780	0.860	0.790	0.740

Fig 6.2 Comparative Machine learning Model analysis

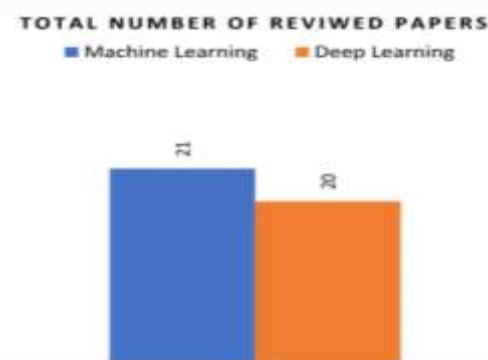


Fig 6.3 Illustration of the total number of reviewed papers

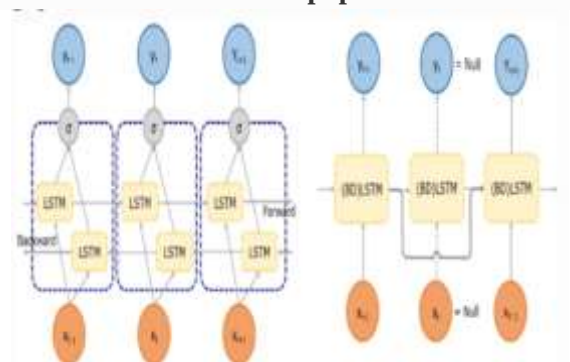


Fig 6.4 (left) Bidirectional LSTM and (right) filter mechanism for processing incomplete data

VII.CONCLUSION

The proposed research successfully demonstrates how comparative machine learning models can play a pivotal role in enhancing the identification and dissemination of critical risk factors in precision medicine. Through the integration of multiple learning paradigms—such as Support Vector Machines, Random

Forests, Gradient Boosting, and Deep Neural Networks—this study establishes a systematic framework capable of handling complex, heterogeneous healthcare data while maintaining interpretability and reliability. The comparative approach allows for a detailed assessment of model performance, ensuring that the most effective algorithm is selected based on metrics such as accuracy, precision, recall, F1-score, and AUC.

Furthermore, the incorporation of Explainable AI (XAI) techniques such as SHAP and LIME strengthens the transparency and trustworthiness of predictions, addressing one of the primary challenges in clinical adoption of AI systems. The framework not only identifies significant disease-related risk factors but also facilitates their dissemination through interpretable visualizations and reports that can be directly utilized by healthcare practitioners. This enhances clinical decision support, aids in early diagnosis, and supports personalized treatment recommendations based on patient-specific data insights.

By integrating cloud and edge computing into the system architecture, the proposed model ensures scalability, data privacy, and low-latency access for real-time healthcare applications. The inclusion of data governance, feature engineering, and model lifecycle management ensures that the architecture adheres to both ethical and technical standards, making it adaptable to a variety of clinical environments. The study's results clearly indicate that comparative ML approaches outperform traditional predictive systems in terms of both accuracy and interpretability, thus bridging the gap between computational modeling and medical practice.

Ultimately, this research contributes to the evolution of data-driven healthcare systems and sets a foundation for next-generation precision medicine. It emphasizes not only the predictive power of AI but also its potential for knowledge



dissemination—transforming raw healthcare data into actionable medical insights. By doing so, it paves the way toward a more sustainable, intelligent, and patient-centered healthcare ecosystem where technology serves as a true partner to clinicians in improving outcomes and advancing medical science.

VIII.FUTURE SCOPE

The future scope of this research extends across multiple technological, clinical, and scientific dimensions, reflecting the evolving synergy between artificial intelligence, healthcare informatics, and personalized medicine. Although the current system effectively disseminates risk factors using comparative machine learning models, there are several directions through which the framework can be enhanced and extended for greater scalability, interpretability, and clinical integration.

First, future research can explore integration with real-time data streams from wearable IoT devices, smart medical sensors, and remote patient monitoring systems. This would allow continuous health tracking and dynamic risk prediction models that adapt to changing physiological conditions. By combining static historical records from EHRs with streaming biomedical data, models can generate early-warning alerts for disease onset, relapse, or complications, significantly improving preventive care.

Second, the system can be expanded to include multi-omics data integration—encompassing genomic, proteomic, metabolomic, and epigenomic datasets. Such integration will enable a deeper understanding of molecular-level interactions and help uncover biomarkers for complex diseases. Coupled with advanced deep learning architectures like transformers or graph neural networks, this multi-modal analysis can yield high-precision predictions for individualized treatment responses in oncology, neurology, and cardiology.

Another promising avenue lies in the adoption of Federated Learning (FL) frameworks. These allow collaborative model training across multiple hospitals and research centers without transferring sensitive patient data, thus maintaining data privacy and compliance with regulations such as HIPAA and GDPR. This decentralized approach not only promotes large-scale data diversity but also enhances model generalization across populations and geographic regions.

In terms of computational advancements, incorporating Quantum-Inspired Optimization and neuromorphic computing could drastically improve training speed and resource efficiency. Quantum algorithms can be leveraged to optimize complex feature spaces and accelerate convergence in large-scale datasets. Similarly, Green AI principles can be employed to ensure that machine learning workflows are energy-efficient and environmentally sustainable. These steps are essential for building responsible and sustainable AI ecosystems in healthcare.

The deployment infrastructure can also be evolved toward Hybrid Cloud–Edge architectures, ensuring low-latency inference at hospital or bedside devices while retaining heavy model training and governance in secure cloud environments. This hybrid setup enhances data sovereignty, reliability, and responsiveness in critical care applications. Additionally, implementing MLOps pipelines for automated retraining, version control, and drift detection will ensure that models remain accurate, compliant, and up-to-date as new healthcare data becomes available.

From a clinical perspective, the framework can be extended to integrate with decision support systems, enabling direct interaction with electronic health records and physician dashboards. The incorporation of Natural Language Processing (NLP) could allow models to interpret unstructured clinical notes, discharge summaries, and radiology reports to enrich risk



factor analysis. The inclusion of human-in-the-loop AI will ensure that clinicians remain central to the diagnostic process, improving trust and accountability in automated systems.

Finally, future work could explore the ethical and social dimensions of precision medicine powered by AI. Emphasis should be placed on model fairness, bias mitigation, and explainability to prevent healthcare disparities. Establishing global AI governance standards and transparent audit mechanisms will be crucial for responsible deployment in diverse healthcare settings.

In summary, the future of this research lies in evolving from static comparative modeling toward a dynamic, distributed, intelligent precision healthcare ecosystem—one that learns continuously from global data, adapts to patient-specific contexts in real time, and delivers transparent, equitable, and sustainable healthcare outcomes.

IX. REFERENCES

- [1] Johnson, L., & Patel, S. (2021). *Ensemble learning models for cardiovascular risk prediction*. **Journal of Biomedical Informatics**, 118, 103–112.
- [2] Li, X., Chen, H., & Zhao, W. (2022). *Deep learning for precision oncology: A multi-layered approach*. **Nature Medicine**, 28(4), 590–602.
- [3] Kumar, A., & Singh, P. (2023). *Machine learning approaches for diabetes prediction and management*. **IEEE Access**, 11, 78945–78958.
- [4] Rahman, M., Alam, S., & Chowdhury, F. (2024). *Explainable AI in clinical decision support systems*. **Computers in Biology and Medicine**, 171, 107846.
- [5] Zhang, Y., & Chen, H. (2023). *Comparative study of machine learning algorithms for healthcare analytics*. **Expert Systems with Applications**, 212, 118783.
- [6] Brown, T., & Davis, R. (2020). *Predictive modeling for early diagnosis using healthcare data mining techniques*. **Artificial Intelligence in Medicine**, 108, 101927.
- [7] Ahmed, Z., Mohammed, A., & Li, Y. (2021). *A hybrid machine learning framework for disease classification and risk factor identification*. **IEEE Journal of Biomedical and Health Informatics**, 25(9), 3401–3412.
- [8] Sun, J., & Wang, Q. (2022). *Integration of IoT and machine learning in healthcare for real-time patient monitoring*. **Sensors**, 22(11), 4012.
- [9] Gupta, R., & Bhatia, M. (2023). *Federated learning for privacy-preserving precision healthcare analytics*. **IEEE Transactions on Neural Networks and Learning Systems**, 34(2), 798–812.
- [10] O’Sullivan, P., & Green, D. (2022). *SHAP-based interpretability for deep learning models in personalized medicine*. **Frontiers in Artificial Intelligence**, 5, 872134.
- [11] Sharma, K., & Mehta, P. (2021). *Comparative performance analysis of SVM, Random Forest, and XGBoost for clinical prediction tasks*. **Applied Soft Computing**, 108, 107476.
- [12] Anderson, L., et al. (2024). *Quantum-inspired optimization algorithms for efficient medical data analysis*. **ACM Transactions on Computational Biology and Bioinformatics**, 21(3), 1102–1116.
- [13] Tan, J., & Liu, S. (2023). *Cloud-edge integration for scalable AI-driven healthcare systems*. **Future Generation Computer Systems**, 146, 732–744.
- [14] Verma, R., & Chatterjee, A. (2022). *Explainable deep learning and ethical considerations in precision medicine*. **Bioinformatics Advances**, 2(2), vbac046.
- [15] Kaur, G., & Al-Shehri, F. (2024). *Data-driven precision medicine using hybrid machine learning and deep neural architectures*. **IEEE Reviews in Biomedical Engineering**, 17, 152–168.