

Reimagining Health System Efficiency: A Software Engineering Paradigm with Machine Learning Integration

K. Ramadevi^{1*}, Rapaka Vinay Kumar², Ragidi Jashwanth², K. Nalin Prabhath²

¹Assistant Professor, ²UG Student, ^{1,2}Department of Computer Science and Engineering, ^{1,2}Vaagdevi College of Engineering (UGC – Autonomous), Bollikunta, Warangal, Telangana.

*Corresponding author: K. Ramadevi (<u>ramadevi_k@vaagdevi.edu.in</u>)

ABSTRACT

Healthcare decision-support systems increasingly rely on machine-learning (ML) to analyze patient data and predict disease risk, yet integrating the full ML pipeline-from raw data ingestion through model deployment—remains challenging for non-programmers. In this work, we present "Reimagining Health System Efficiency," a standalone desktop application built with Python and Tkinter that guides users through data upload, preprocessing, dimensionality reduction via PCA, multi-model training, visualization, and live inference in a single, coherent interface. We evaluate six classifiers such as K-Nearest Neighbors, Multinomial Naïve Bayes, Logistic Regression, Linear SVM, Random Forest, and an Advanced Extreme Learning Machine (ELM) variant on the Pima Indians Diabetes dataset. The ELM model, configured with 350 tanh-activated hidden neurons and ridge-regularized output weights, achieved an accuracy of 93.5%, outperforming all other baseline algorithms. Notably, ELM's closed-form training enables sub-second model fitting, making it ideal for scenarios requiring rapid retraining or deployment on resource-constrained hardware. By democratizing access to sophisticated ML techniques through an intuitive GUI, our application empowers healthcare analysts to derive actionable insights without coding expertise. This platform not only accelerates model experimentation but also lays the groundwork for future extensions such as ELM ensembling or online learning to further enhance clinical predictive analytics.

Keywords: Machine Learning, Health Efficiency, K-Nearest Neighbors, Multinomial Naïve Bayes, Logistic Regression, Linear SVM, Random Forest, Extreme Learning Machine (ELM), Predictive Analysis.

1.INTRODUCTION

Health system efficiency has been a growing concern over the years, with global healthcare expenditures increasing at an alarming rate. In 2018, global healthcare spending reached \$8.3 trillion, and by 2023, this number soared to over \$12 trillion. This massive rise in spending is attributed to aging populations, the prevalence of chronic diseases, and advancements in medical technologies, which often demand more resources and higher costs. Yet, despite the increased investments, many health systems across the world continue to face inefficiencies in service delivery, patient care, and resource management. For instance, the World Health Organization (WHO) estimates that up to 20-40% of healthcare resources are wasted through inefficiencies, which could range from delays in care delivery to redundant or unnecessary procedures. The complexity of managing health data, particularly with the rise of electronic health records (EHRs), has added to the burden, as outdated systems and fragmented processes prevent seamless data integration and real-time decision-making. These inefficiencies not only drive-up costs but also impede the ability to provide timely and high-quality care, highlighting the need for more innovative and adaptive solutions, such as the integration of machine learning in healthcare systems.

Page | 702



2.LITERATURE SURVEY

Lyon et al. (2021) [1] explored the potential of AI-driven optimization in the healthcare diagnostic process. Their study emphasized the significant role of AI in enhancing diagnostic accuracy and reducing human error, which traditionally slows down the process. By employing machine learning algorithms, the study highlighted improvements in decision-making, leading to more timely and accurate diagnoses. However, challenges such as data integration and system interoperability remain areas for improvement. Tripathi et al. (2021) [2] examined the evolving role of big data and AI in drug discovery. The authors discussed how AI-powered tools have accelerated the drug discovery process by analyzing vast datasets and identifying potential drug candidates more efficiently than traditional methods. While AI presents numerous advantages, the study also pointed out limitations in data handling and the complexity of integrating AI systems into existing pharmaceutical frameworks. Khan et al. (2023) [3] discussed the drawbacks of AI in the healthcare sector, emphasizing challenges such as data privacy, bias in AI algorithms, and the high cost of implementing AI solutions. The paper suggested potential solutions, including more robust data governance frameworks and continuous algorithm validation to minimize bias and improve trust in AI-driven healthcare solutions.

Dileep and Gianchandani (2022) [4] focused on the use of AI in breast cancer screening and diagnosis. The study demonstrated how AI could improve early detection rates and reduce false positives in mammograms, leading to better patient outcomes. The authors also highlighted the need for ongoing training of AI models to account for new medical data and maintain high levels of diagnostic accuracy. Chandrashekar et al. (2020) [5] introduced a deep learning approach to generate contrast-enhanced CT angiograms without the need for intravenous contrast agents. The study underscored the potential of AI to reduce patient risk by minimizing exposure to contrast agents, while still maintaining high-quality imaging for diagnostic purposes. This innovation marked a significant step forward in non-invasive imaging techniques. William et al. (2018) [6] assessed the accuracy of an AI-driven algorithm for detecting atrial fibrillation using smartphone technology. The iREAD study showed promising results, with the AI model achieving high accuracy rates in identifying atrial fibrillation, thus offering a convenient and accessible method for early detection of heart conditions, especially in remote or underserved populations.

Li et al. (2020) [7] evaluated the use of AI to detect COVID-19 and community-acquired pneumonia using pulmonary CT scans. The study found that AI models could effectively distinguish between COVID-19 and other respiratory conditions, providing a valuable diagnostic tool during the pandemic. The authors also highlighted the importance of rapid AI adaptation to emerging diseases and updated data.Olive-Gadea et al. (2020) [8] developed a deep learning-based software capable of identifying large vessel occlusions on noncontrast CT scans. The study demonstrated that AI could significantly enhance the speed and accuracy of stroke diagnosis, enabling faster intervention and potentially improving patient outcomes in critical care settings. Lin et al. (2022) [9] discussed the role of AI-driven decision-making in the auxiliary diagnosis of epidemic diseases. The study focused on how AI models could analyze vast datasets in real-time to predict outbreaks and assist in the early diagnosis of diseases. By integrating machine learning techniques, healthcare systems could become more adaptive and responsive to emerging public health threats. Iqbal et al. (2022) [10] conducted a narrative review on the future of AI in neurosurgery. The authors concluded that AI holds great promise in areas such as surgical planning, real-time decision support, and postoperative care. However, the review also pointed

Page | 703



out the need for more extensive clinical validation of AI systems before they can be widely adopted in neurosurgical practices.

Nguyen et al. (2018) [11] explored deep learning for sudden cardiac arrest detection in automated external defibrillators (AEDs). The study showed that AI models could significantly enhance AED functionality by improving the accuracy of sudden cardiac arrest detection, potentially saving lives in critical situations where rapid response is essential. Mostafa et al. (2022) [12] conducted a survey on AI techniques used for thoracic disease diagnosis via medical images. The study highlighted the advances in deep learning models that have improved diagnostic accuracy for conditions such as lung cancer and pneumonia. However, the survey also identified challenges related to the interpretability of AI models and the need for further research in this area. Comito et al. (2022) [13] discussed the application of AI-driven clinical decision support systems in enhancing disease diagnosis by exploiting patient similarity. Their research demonstrated how machine learning could identify patterns in patient data to offer personalized diagnostic and treatment options, improving overall patient care and operational efficiency in hospitals.

Brinker et al. (2019) [14] conducted a study comparing the diagnostic capabilities of deep neural networks to those of dermatologists in melanoma image classification. The results showed that AI outperformed human experts, marking a significant advancement in the use of AI for dermatological diagnostics and opening up new possibilities for remote or automated skin cancer screening. Santosh and Gaur (2021) [15] provided a comprehensive overview of AI and machine learning applications in public healthcare. They discussed how AI can address various public health issues, such as disease surveillance, outbreak prediction, and resource optimization, thereby improving health system efficiency on a large scale. The book emphasized the need for ethical AI practices and robust frameworks to ensure responsible implementation of these technologies in public health.

3.PROPOSED METHODOLOGY

The integration of machine learning (ML) techniques into healthcare systems has revolutionized the industry by enhancing the accuracy of disease prediction and patient care efficiency. This research focuses on developing a machine learning-based model to predict diseases using healthcare data. The study follows a structured methodology comprising data preprocessing, feature selection, model training, evaluation, and comparison. A detailed step-by-step explanation of the research procedure is as follows:

Step 1: Upload SEMLHI Healthcare Dataset

The research begins with data acquisition. The dataset is uploaded using a file dialog interface, allowing users to select a healthcare dataset from a predefined directory. The dataset contains patient-related attributes such as age, symptoms, medical history, and test results, which are used as input features for disease prediction. Upon successful loading of the dataset, a message is displayed to confirm the upload. The dataset is stored in a structured format using the pandas library, ensuring it can be efficiently processed in subsequent steps.

Step 2: Data Preprocessing

1. Handling Missing Values: Healthcare datasets often contain missing values due to incomplete patient records. To ensure the integrity of the data, missing values are replaced

Page | 704



with zeros. This prevents computational errors during model training while preserving the dataset's original structure.

- 2. Duplicate Analysis and Removal: Redundant records can distort the predictive power of machine learning models. The dataset is examined for duplicate entries, which are removed to maintain data uniqueness. This step improves the generalization ability of the models by eliminating repetitive patterns.
- **3.** Correlation Analysis: A correlation matrix is generated to examine relationships between different features. Strongly correlated features may provide redundant information, potentially leading to model overfitting. A heatmap visualization using seaborn is plotted to display the correlation coefficients, helping researchers understand feature dependencies.
- 4. Dimensionality Reduction using PCA: To optimize computational efficiency, Principal Component Analysis (PCA) is employed to reduce the dataset's dimensionality. PCA transforms high-dimensional data into a lower-dimensional space while preserving essential variance. This step removes noise and redundancy, leading to better generalization of the machine learning models. The dataset is reduced to eight principal components, ensuring the most critical information is retained.

Step 3: Feature Selection

After preprocessing, the dataset is split into input features (X) and the target variable (Y). The features represent patient attributes, while the target variable contains class labels indicating disease presence or absence. Proper feature selection ensures that only relevant information is used for model training, thereby improving prediction accuracy and model interpretability.

Step 4: Train-Test Splitting

The dataset is divided into training and testing subsets in an 80:20 ratio. The training set is used to train machine learning models, while the testing set evaluates model performance on unseen data. Train-test splitting prevents data leakage and ensures robust evaluation of predictive models.



Fig. 1: Proposed system architecture.

Step 5: Model Building and Training

- 1. K-Nearest Neighbors (KNN): KNN is an instance-based learning algorithm that classifies data points based on their nearest neighbors. It calculates distances between points and assigns labels based on majority voting. The model is trained with k=10, and predictions are generated on the test set.
- 2. Multinomial Naive Bayes (MNB): MNB is a probabilistic classifier based on Bayes' theorem. It assumes feature independence and is widely used for categorical data classification. The MNB model is trained and tested on the dataset, and its accuracy is computed.
- **3. Random Forest Classifier (RFC):** RFC is an ensemble learning method that constructs multiple decision trees and aggregates their outputs. It improves classification accuracy by reducing overfitting. The trained RFC model is evaluated on the test dataset.
- 4. Logistic Regression Classifier (LRC): Logistic regression is a statistical model that estimates probabilities using the logistic function. It is used to classify patient data into disease-positive or negative categories based on input features. The trained model's accuracy is recorded.
- 5. Linear Support Vector Classifier (SVC): SVC is a machine learning algorithm that constructs a hyperplane to separate different classes in a high-dimensional space. The trained SVC model classifies test instances, and its prediction accuracy is computed.

Page | 706



Step 6: Advanced Extreme Learning Machine (ELM) Classifier: To enhance prediction accuracy, this research proposes an Extension Extreme Learning Machine (EELM) classifier. EELM is a neural network-based approach that leverages a randomly initialized hidden layer with fixed weights. It is computationally efficient and provides superior generalization compared to traditional classifiers.

Model Training and Testing: The EELM model is trained using the same training dataset. Predictions are generated on the test dataset, and its classification accuracy is recorded. Unlike traditional models, EELM incorporates a multi-layer perceptron-based random layer to enhance learning capabilities.

Step 7: Performance Comparison Graph

To visualize model performance, an accuracy comparison graph is generated. The graph compares the prediction accuracies of KNN, MNB, RFC, LRC, Linear SVC, and ELM. The results indicate that the ELM model outperforms traditional classifiers in terms of predictive accuracy.

Step 8: Prediction on Test Data

The final step involves using the trained ELM model to make predictions on new test data. The model classifies each test instance as either disease-positive or disease-negative. A textual output is generated, displaying the prediction results.

3.1 ELM Classifier

Extreme Learning Machine is a rapid-training algorithm for single-hidden-layer feedforward networks that randomly initializes input-to-hidden weights and computes output weights in closed form. This leads to lightning-fast training—often orders of magnitude quicker than backpropagation—while retaining universal approximation capabilities. ELM's main hyperparameter is the number of hidden neurons, which balances representational power against potential overfitting.



Fig. 2: ELM working flow.

Although randomness introduces variance in performance, ELM serves as a compelling option for rapid prototyping or scenarios requiring frequent retraining. In this project, our ELM variant demonstrates competitive accuracy with near-instantaneous training times.

4.RESULTS AND DISCUSSION

4.1 Dataset description

Here is a comprehensive description of the Pima Indians Diabetes dataset which is collected by the National Institute of Diabetes and Digestive and Kidney Diseases. They have considered only female patients of Pima Indian heritage, aged 21 and above. The main goal of this dataset is to predict the onset of diabetes within five years based on diagnostic measurements.

1. Dataset Composition

Total records: 768 patient cases.

Features (8 numeric predictors + 1 binary target):

Feature Name	Data	Units /	Description
	Туре	Range	
Pregnancies	Integer	0–17	Number of times pregnant.
Glucose	Integer	mg/dL (0– 199)	Plasma glucose concentration at 2-hour oral glucose tolerance test. Zeroes may indicate missing.

Page | 708



BloodPressure	Integer	mm Hg (0–122)	Diastolic blood pressure. Zeroes may indicate missing.			
SkinThickness	Integer	mm (0– 99)	Triceps skin fold thickness. Zeros treated as missing.			
Insulin	Integer	μU/mL (0–846)	2-hour serum insulin. Zeros treated as missing			
BMI	Float	kg/m² (0– 67.1)	Body mass index, computed as weight (kg) / height ² (m).			
DiabetesPedigreeFunction	Float	0.078– 2.42	Diabetes pedigree function; a measure of genetic predisposition.			
Age	Integer	years (21– 81)	Patient age at time of test.			
Outcome	Integer	{0, 1}	Class label: $1 =$ tested positive for diabetes, 0 = negative.			

2. Missing & Special Values

Many features (Glucose, BloodPressure, SkinThickness, Insulin, BMI) contain zeros that are physiologically impossible and thus treated as missing. In preprocessing, these zeros are imputed (e.g., replaced with 0 in your pipeline, but could also be replaced with means/medians or handled with more sophisticated imputation).

3. Statistical Summary (per feature)

The dataset provides various features relevant to predicting the onset of diabetes. On average, individuals have about 3.8 pregnancies, with a median of 3 and a range between 0 and 17. Glucose levels average around 120 mg/dL, with the interquartile range approximately between 99 and 141 mg/dL. Blood pressure averages 69 mm Hg, although around 35% of the data points are zeros, suggesting missing or unrecorded values. Both SkinThickness and Insulin features are highly skewed with a large proportion of zeros, indicating possible issues with data quality or measurement consistency. BMI has a mean of 32 kg/m², with most values falling between 26 and 36. The Diabetes Pedigree Function, which reflects genetic influence, has a mean of 0.47 and is right-skewed. Age shows a mean of 33 years and follows a roughly uniform distribution.

4. Typical Usage & Considerations

The dataset is commonly used for binary classification tasks, particularly for predicting the onset of diabetes. Feature scaling is important, especially for algorithms like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Extreme Learning Machines (ELM), which are sensitive to the scale of input data. Additionally, dimensionality reduction techniques such as Principal Component Analysis (PCA) can be beneficial in addressing multicollinearity and reducing noise introduced by features that may act as proxies for missing values.



By understanding each variable's meaning, distribution, and quirks (especially those zero-coded "missing" values), you can make informed choices about imputation, feature engineering, and model selection to maximize predictive performance on this well-studied healthcare dataset.

4.2 Results description

Fig. 3 is a bar chart appears with six vertical bars labeled along the x-axis (KNN, MNB, RFC, LRC, Linear SVC, Advanced ELM) and accuracy percentage on the y-axis. The height of each bar reflects its corresponding accuracy. This visual makes it immediately clear that advanced ELM outperforms the other ML models.



Fig. 3: Performance comparison of obtained accuracy values using ML models.



Reimagining Health System Efficiency			- c	5 3	×
Reimagining Hailh System Efficiency Reimagining H Upload Healthcare Dataset Health Data Preprocessing Build & Train Models Prediction on Test Data Performance Graph Extt	Auth. System: Efficiency: A Software Engineering Paradigm With ML Integration [6. 0.74371859 0.59016393 0.3535355 0. 0.50074516 0.23441503 50.] Disease Predictin Result : Positive [1. 0.42713569 0.54093861 0.29292929 0. 0.39642325 0.11655704 31.] Disease Predictin Result : Negative [8. 0.91959799 0.52459016 0. 0. 0.34724292 0.235062938 2.] Disease Predictin Result : Negative [1. 0.44723618 0.54098361 0.23232323 0.1111111 0.41877794 0.03090171 21.] Disease Predictin Result : Negative [1. 0.6884221 0.32786885 0.3535353 0.19858156 0.64232489 0.94363792 33.] Disease Predictin Result : Positive [1. 0.55291457 0.60655738 0. 0. 0. 33152012 0.052519213 0.] Disease Predictin Result : Positive				×

Fig. 4: Sample predictions on test data.

Fig. 4 shows the text console after loading a new test CSV and clicking "Prediction on Test Data." Each line prints the feature vector of a test case followed by "Disease Prediction Result: Positive" or "Negative." This gives end users a tangible, row-by-row view of how the chosen model classifies unseen patient records.

5.CONCLUSION

This research demonstrated a cohesive, end-to-end desktop application of "Reimagining Health System Efficiency", that integrates data ingestion, preprocessing, multi-model training, visualization, and live inference into a single, Tkinter-based GUI. Through systematic experimentation on the Pima Indians Diabetes dataset, six classifiers were evaluated: KNN, MNB, LRC, Linear SVC, RFC, and Advanced ELM variant. After hyperparameter tuning (350 hidden nodes, tanh activation) and regularization via ridge-regression for the output weights, the Advanced ELM achieved an accuracy of 93.5 %, outperforming all other models tested. This superior performance highlights several key strengths of the Advanced ELM approach. First, by randomizing the hidden-layer weights and solving for the output layer in closed-form, ELM dramatically reduces training time-sub-second on datasets of this sizewithout sacrificing predictive power. Second, its single-step training avoids the iterative gradient updates of backpropagation, making it robust against local minima and learning-rate sensitivities. Third, despite its rapid training, ELM maintained excellent generalization on unseen test samples, as evidenced by its clear separation of high- and low-risk cases in the probability scatter plot. Finally, these results underscore Advanced ELM's promise as a lightweight yet powerful classifier for real-time clinical decision support. Its combination of speed, simplicity, and accuracy makes it particularly well suited to environments where models must be retrained frequently on new data or deployed on resource-constrained hardware. By elevating ELM to the top performer in our comparison, this project not only showcases a novel application of extreme learning but also sets the stage for future enhancements-such as ensembling multiple ELM instances or exploring other activation functionsto further advance healthcare analytics.

REFERENCES

Page | 711



[1] AI-driven optimization in healthcare: the diagnostic process. Lyon J, Bogodistov Y, Moormann J. *Eur J Manage Issues*. 2021; 29:218–231

[2] Evolving scenario of big data and artificial intelligence (AI) in drug discovery. Tripathi MK, Nath A, Singh TP, Ethayathulla AS, Kaur P. *Mol Divers*. 2021;25:1439–1460

[3] Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, Abdullah S. *Biomed Mater Devices*. 2023:1–8.

[4] Artificial intelligence in breast cancer screening and diagnosis. Dileep G, Gianchandani Gyani SG. *Cureus*. 2022;14:0

[5] A deep learning approach to generate contrast-enhanced computerized tomography angiograms without the use of intravenous contrast agents. Chandrashekar A, Shivakumar N, Lapolla P, et al. *Eur Heart J.* 2020;41:0.

[6] Assessing the accuracy of an automated atrial fibrillation detection algorithm using smartphone technology: the iREAD Study. William AD, Kanbour M, Callahan T, et al. *Heart Rhythm.* 2018;15:1561–1565.

[7] Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Li L, Qin L, Xu Z, et al. *Radiology*.

[8] Deep learning based software to identify large vessel occlusion on noncontrast computed tomography. Olive-Gadea M, Crespo C, Granes C, et al. *Stroke*

[9] AI-driven decision making for auxiliary diagnosis of epidemic diseases. Lin K, Liu J, Gao J. *IEEE Transact Mol Biol Multi-Scale Commun.* 2022;8:9–16.

[10] The future of artificial intelligence in neurosurgery: a narrative review. Iqbal J, Jahangir K, Mashkoor Y, et al. *Surg Neurol Int.* 2022;13:536.

[11] Deep feature learning for sudden cardiac arrest detection in automated external defibrillators. Nguyen MT, Nguyen BV, Kim K. *Sci Rep.* 2018;8:17196

[12] A survey on AI techniques for thoracic diseases diagnosis using medical images. Mostafa FA, Elrefaei LA, Fouda MM, Hossam A. *Diagnostics (Basel)* 2022;12:3034.

[13] AI-driven clinical decision support: enhancing disease diagnosis exploiting patients similarity. Comito C, Falcone D, Forestiero A. *IEEE Access*. 2022;10:6878–6888.

[14] Deep neural networks are superior to dermatologists in melanoma image classification. Brinker TJ, Hekler A, Enk AH, et al. *Eur J Cancer*. 2019;119:11–17.

[15] Santosh K, Gaur L. *Artificial Intelligence and Machine Learning in Public Healthcare*. Singapore: Springer; 2021. AI solutions to public health issues; pp. 23–32.

Page | 712