SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Hybrid Machine Learning for Healthcare Disease Prediction Using Latent Feature and Anomaly Learning

R. Kala[1], Dr. K. K. Savitha[2]

[1]Research Scholar, Bharathiar University PG Extension and Research Centre, Perundurai, Erode, India. E-mail: kalamca05@gmail.com
[2]Assistant Professor, Bharathiar University PG Extension and Research Centre, Perundurai, Erode, India. E-mail: savitha.gopinath@gmail.com

Abstract
Anomaly detection in healthcare systems is one way through which disease prediction can be made accurate and the reliability of the data utilized in making clinical decisions ensured. The use of traditional machine learning algorithms may not be able to detect anomalies, nonlinear and complex patterns in the healthcare data. This study aims to design an anomaly aware healthcare prediction model using the integration of an autoencoder for latent feature extraction and isolation forest, K-means clustering and gradient-boosted classification for disease prediction. This model was tested using two healthcare datasets that contained 2,000 diabetes data records and 50,000 cancer data records. Various models were designed and analyzed to find out how useful the inclusion of anomaly detection, latent feature extraction, and classification is in disease prediction. Of all the tested models, the Auto Encoder (AE) + Isolation Forest + LightGBM combination provided the best performance on the diabetes dataset with an accuracy of 90.75% and ROC-AUC value of 0.9593, while Auto Encoder (AE) + K-Means + LightGBM combination provided the best performance on the cancer dataset with an accuracy of 97.51% and ROC-AUC value of 0.9969. From the above results, the paper concludes that using Autoencoders to extract latent space features along with LightGBM works well for anomaly-aware disease prediction.

Keywords: Healthcare Anomaly Detection, Autoencoder, K-Means Clustering, LightGBM , Disease Prediction, Hybrid Machine Learning

Introduction

Healthcare data analytics has become an essential component of modern medical diagnosis and decision support systems due to the increasing availability of large-scale electronic health records and medical datasets. Diseases such as diabetes and cancer continue to represent significant global health challenges because of their high prevalence, mortality rates, and complex clinical characteristics [1]. Early identification of abnormal patient conditions and reliable disease prediction are therefore critical for improving treatment outcomes and reducing healthcare costs.

Traditional machine learning algorithms have demonstrated considerable success in disease classification tasks; however, they often focus solely on predictive accuracy while overlooking hidden anomalies and irregular patterns present in healthcare datasets [2]. Medical datasets frequently contain missing values, corrupted entries, inconsistent records, outliers, and overlapping class distributions that can negatively affect prediction reliability and clinical decision-making processes. Furthermore, healthcare data are typically high-dimensional and exhibit nonlinear relationships among variables, making accurate disease prediction a challenging task for conventional models [3].

Anomaly detection techniques such as Isolation Forest and clustering approaches such as K-Means have been widely used to identify unusual patterns and abnormal observations in data. Although these unsupervised methods are useful for detecting anomalies, they often struggle to establish accurate decision boundaries required for disease classification. Similarly, clustering algorithms may fail to capture the complex interactions among healthcare variables, particularly in datasets with high dimensionality and heterogeneous characteristics.

Recent advances in deep learning have introduced Autoencoders as an effective mechanism for extracting compact latent representations from complex datasets. Autoencoders are capable of learning hidden feature structures while reducing dimensionality and preserving essential information from the original data [4]. These learned representations can provide a stronger basis for downstream classification tasks compared to raw input features or anomaly scores alone.

Motivated by these limitations, this study proposes a hybrid anomaly-aware healthcare prediction framework that integrates Autoencoder-based feature extraction with Isolation Forest, K-Means clustering, and LightGBM classification. The proposed approach was evaluated using diabetes and cancer datasets containing 2,000 and 50,000 records respectively. Six model combinations were systematically compared to investigate the contribution of anomaly detection, clustering, and supervised boosting techniques in healthcare prediction [5].

Experimental findings demonstrate that standalone unsupervised methods exhibit weak predictive capability, whereas Autoencoder-based hybrid architectures integrated with LightGBM significantly improve classification accuracy and Area Under Curve (AUC) performance [6-10]. Among all evaluated methods, the Autoencoder + K-Means + LightGBM framework achieved the highest overall accuracy and consistency across both healthcare datasets, making it a promising solution for future real-time healthcare anomaly detection applications.

- A comparative hybrid healthcare anomaly detection framework integrating Autoencoder, Isolation Forest, K-Means, and LightGBM was developed and evaluated on diabetes and cancer datasets.
- Six different model combinations were systematically analyzed to investigate the effectiveness of unsupervised anomaly detection and supervised classification approaches for healthcare prediction.
- The study demonstrated that Autoencoder-based latent feature extraction combined with LightGBM significantly improves classification accuracy and AUC, with AE + KMeans + LightGBM emerging as the best-performing model.

The remainder of this paper is organized as follows. The related work section reviews existing studies on healthcare anomaly detection, Autoencoder-based feature learning, clustering techniques, and boosting algorithms for medical prediction. The materials and methods section describe the datasets, pre-processing techniques, Autoencoder architecture, and hybrid model construction process. The results and discussion section present comparative performance analysis using Accuracy, Precision, Recall, AUC, and ROC metrics. Finally, the conclusion summarizes the major findings and outlines future directions toward real-time and cloud-enabled healthcare anomaly detection systems.

2. Related Work

The rapid growth of healthcare data has encouraged the development of machine learning and deep learning techniques for disease diagnosis, anomaly detection, and predictive analytics. Traditional machine learning frameworks have achieved considerable success in healthcare prediction tasks; however, the increasing complexity and dimensionality of medical datasets demand more advanced analytical approaches. [11] introduced the Scikit-learn framework, which became one of the most widely used machine learning libraries for implementing classification, clustering, and anomaly detection algorithms in healthcare analytics. The library provides efficient implementations of algorithms such as K-Means clustering and Isolation Forest, facilitating reproducible medical data analysis. Anomaly detection has gained significant attention due to its capability to identify abnormal patient records and hidden disease patterns. [12] proposed anomaly detection and explanation mechanisms for healthcare data, highlighting the importance of interpretability in clinical decision support systems. Similarly, [13] improved the conventional Isolation Forest algorithm using attention mechanisms to enhance anomaly identification in high-dimensional datasets.

Deep anomaly detection techniques have further improved healthcare analytics. [14] introduced Deep Isolation Forest, which combines representation learning with anomaly detection to improve detection accuracy in complex datasets. [15] demonstrated the effectiveness of Isolation Forest in healthcare monitoring applications by detecting anomalies in patient vital signs and physiological measurements. Ensemble learning approaches have also contributed significantly to healthcare prediction systems.[16] discussed the effectiveness of ensemble learning methods in improving model robustness and generalization performance. [17] further emphasized the advantages of combining multiple learning algorithms for handling nonlinear and high-dimensional datasets. Recent advances in deep learning have expanded the application of neural networks in biomedical research. [18] demonstrated the potential of deep learning models in genomic prediction and healthcare analytics, while [19] highlighted the transformative role of artificial intelligence in medical imaging and radiology diagnostics. Furthermore, [20] proposed foundation models for genomics, demonstrating the ability of large-scale neural architectures to extract meaningful biological representations from complex datasets. Although previous studies have

explored anomaly detection, deep learning, and ensemble methods individually, limited research has investigated the integration of Autoencoder-based latent feature extraction with anomaly detection and supervised boosting methods for disease prediction. This motivates the development of the proposed hybrid framework combining Autoencoder, Isolation Forest, K-Means clustering, and XGBoost for healthcare anomaly detection and classification.

Ref	Techniques Used	Outcome Metrics	Advantages	Limitations
[11]	Scikit-learn ML Framework	Accuracy, Precision	Easy implementation of ML algorithms	No healthcare-specific optimization
[12]	Healthcare anomaly	Accuracy, Explainability	Improved anomaly	Limited deep feature

	detection		interpretation	extraction
[13]	Attention-based Isolation Forest	Detection Accuracy	Better anomaly detection capability	Increased computational complexity
[14]	Deep Isolation Forest	AUC, Accuracy	Handles complex feature spaces	Lack of supervised learning integration
[15]	Isolation Forest	Precision, Recall	Effective vital sign monitoring	Limited disease classification ability
[16]	Ensemble Learning	Accuracy	Improved model robustness	Requires careful model selection
[17]	Machine Learning Frameworks	Generalization Accuracy	Handles nonlinear data	Sensitive to feature quality
[18]	Deep Learning in Genomics	Prediction Accuracy	Learns complex representations	Requires large training datasets
[19]	AI in Medical Imaging	Diagnostic Accuracy	Supports radiological diagnosis	Limited anomaly detection focus
[20]	Foundation Models for Genomics	Representation Quality	Strong feature extraction capability	High computational requirements

2.1 Research Gap

Existing studies primarily investigate anomaly detection, clustering, deep learning, and ensemble learning techniques independently for healthcare analytics. Isolation Forest and clustering methods often fail to capture complex disease-specific decision boundaries, while deep learning models alone may overlook anomaly information. Furthermore, limited studies [21-25] have evaluated hybrid frameworks combining Autoencoder-based latent feature extraction with both anomaly detection and supervised boosting algorithms across multiple healthcare datasets. Therefore, there remains a need for an integrated anomaly-aware healthcare prediction framework capable of improving classification accuracy, robustness, and reliability in high-dimensional medical datasets.

3. Methodology

The proposed study develops a hybrid anomaly-aware healthcare prediction framework integrating Autoencoder, Isolation Forest, K-Means clustering, XGBoost and LightGBM classification techniques for disease prediction and anomaly detection. The methodology consists of healthcare dataset acquisition, preprocessing, latent feature extraction, anomaly analysis, and supervised classification stages. Initially, diabetes and cancer datasets are collected and prepared for model training. Data cleaning, normalization, and dataset partitioning are then performed to improve data quality and consistency. Subsequently, Autoencoder networks generate compressed latent representations of healthcare records. These representations are utilized by Isolation Forest and K-Means algorithms to detect anomalies and hidden structures. Finally, LightGBM combines latent features and anomaly information to perform disease classification.

3.1 Healthcare Dataset Collection and Preparation

The proposed framework employs two healthcare datasets containing diabetes and cancer patient records for anomaly detection and disease prediction. The diabetes dataset contains 2,000 observations with eight clinical attributes, whereas the cancer dataset consists of 50,000 instances containing thirty diagnostic measurements. These datasets provide heterogeneous and high-dimensional characteristics suitable for evaluating hybrid machine learning models. Prior to model development, the datasets undergo preprocessing procedures including missing value treatment, noise removal, outlier elimination, and feature normalization to improve data consistency and analytical performance. The preprocessing stage ensures that the learning algorithms receive standardized input features and reduces the influence of data imbalance and measurement variability.

The feature matrix of the healthcare dataset is represented as in eqn 1:

$$X = \{x_1, x_2, .., x_n\} \quad (1)$$

where, X = Healthcare feature dataset, xi = Individual patient feature vector, n = Total number of patient records
The class label vector is defined as in eqn 2:

$$Y = \{y_1, y_2, .., y_n\} \quad (2)$$

where, Y = Target output vector, yi = Disease class label for patient i, n = Number of observations
The standardized datasets are subsequently supplied to the Autoencoder and Random Forest models for latent feature extraction, anomaly detection, and disease classification. The collected datasets contain numerical clinical measurements and diagnostic information that facilitate disease prediction and anomaly identification under different healthcare scenarios.

3.1.1 Diabetes Dataset

The diabetes dataset contains clinical variables including pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree function, and patient age. The target variable indicates whether a patient is diabetic or non-diabetic based on medical diagnosis. These clinical indicators are extensively utilized in diabetes risk assessment and predictive healthcare analytics because they represent important metabolic and physiological characteristics associated with disease progression. The dataset contains moderate dimensionality and structured numerical information, making it highly suitable for evaluating anomaly-aware machine learning algorithms and latent feature extraction techniques. Interdependencies among glucose concentration, insulin response, and body mass index significantly influence disease prediction accuracy and contribute to the learning capability of the Autoencoder model.
The glucose normalization process is expressed as in eqn 3:

$$G_n = \frac{G_i - G_{min}}{G_{max} - G_{min}} \quad (3)$$

where, Gn = Normalized glucose value, Gi = Original glucose measurement, Gmin= Minimum glucose value in the dataset, Gmax = Maximum glucose value in the dataset
The diabetes classification label is represented as in eqn 4:

$$D_i = \{1, \ if \ diabetics \ is \ present \ 0, otherwise \quad (4)$$

where, Di = Diabetes diagnosis label for patient i
The preprocessing stage removes missing values and outliers to improve data quality and enhance model robustness. Feature scaling ensures that all clinical variables contribute equally during model training and optimization. The Autoencoder learns compressed latent representations from these normalized attributes, enabling efficient reconstruction of patient profiles and identification of abnormal health patterns. Such latent representations improve anomaly detection capability and support accurate diabetes prediction under varying clinical conditions. The normalized clinical measurements further improve model convergence and facilitate reliable anomaly detection in diabetic patient records.

3.1.2 Cancer Dataset

The cancer dataset contains thirty diagnostic features obtained from medical examinations, histopathological analysis, and imaging measurements. The target variable classifies tumors as malignant

or benign according to clinical diagnosis and laboratory findings. Important attributes include radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension measurements that characterize tumor morphology and structural properties. Due to the presence of complex nonlinear relationships among these features, the dataset provides an appropriate benchmark for evaluating deep learning-based feature extraction and anomaly detection approaches. The high dimensionality of the dataset allows the proposed hybrid framework to identify subtle variations between malignant and benign samples that may not be easily distinguishable using conventional machine learning methods.
The tumor feature vector is defined as in eqn 5:

$$C_i = \{C_1, C_2, \dots, C_n\} \quad (5)$$

where,Ci = Feature vector of tumor sample i Cm = Diagnostic feature value
m = Total number of diagnostic attributes
The cancer classification function is represented as in eqn 6:

$$T_i = \{1, \ if \ tumor \ is \ malignant \ 0, if \ tumor \ is \ benign \quad (6)$$

where, Ti = Tumor diagnosis label for sample i
Before model training, the diagnostic attributes undergo preprocessing operations including normalization and feature standardization to eliminate scale variations among measurements. The Autoencoder extracts

meaningful latent representations that preserve important structural information associated with tumor characteristics. These compressed features reduce dimensionality while retaining discriminative information required for classification. Consequently, the hybrid learning framework effectively identifies anomalous samples and improves differentiation between malignant and benign tumor categories in complex healthcare environments. The diagnostic measurements and class labels therefore provide a robust environment for evaluating anomaly detection and disease prediction performance in healthcare applications.

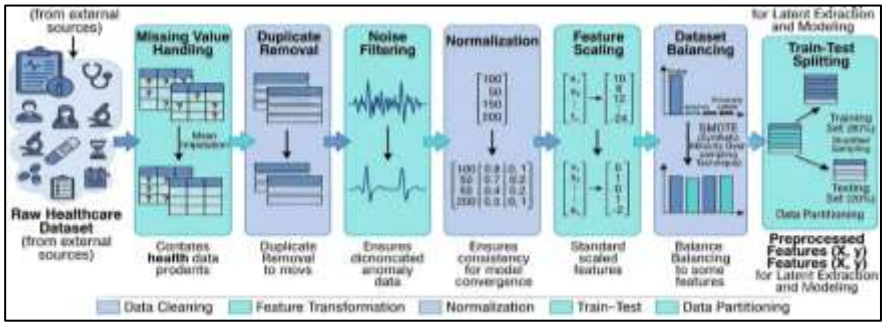
3.2 Data Preprocessing and Feature Engineering

Data preprocessing was performed to improve the quality, consistency, and reliability of healthcare records prior to model development. Missing values, duplicate observations, and noisy entries were identified and corrected to minimize prediction bias and improve dataset integrity. Outlier detection techniques were employed to remove abnormal observations that could adversely influence model performance. Feature scaling and normalization were applied to maintain uniform value ranges among clinical variables and prevent attributes with larger magnitudes from dominating the learning process. Furthermore, feature engineering techniques were utilized to extract meaningful representations and enhance the discriminative capability of the predictive model.

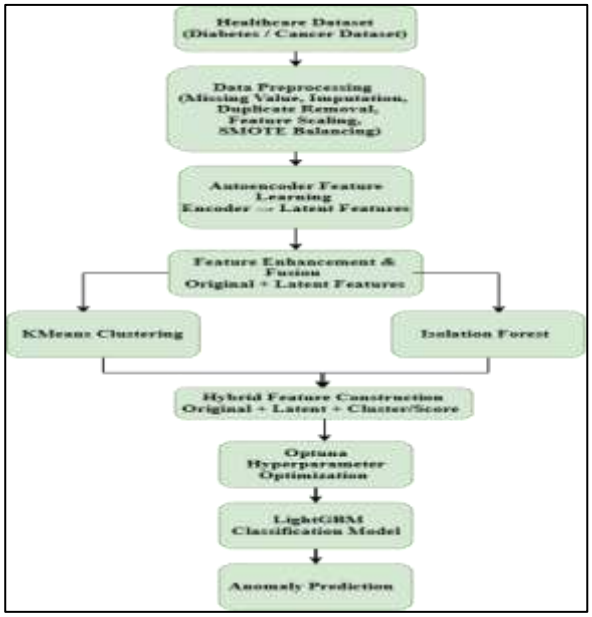
The normalized feature value is calculated as in eqn 7:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (7)$$

where, Xnorm = Normalized feature value, X = Original feature value, Xmin = Minimum feature value, Xmax = Maximum feature value



(a)



(b)

Figure 1: (a) Healthcare Data Pre-processing and Feature Engineering Pipeline, (b) Flow of the proposed methodology

Figure 1 presents the healthcare data preprocessing pipeline adopted in this study. The process includes missing value treatment, duplicate removal, normalization, feature scaling, and stratified dataset partitioning. These operations improve feature consistency and reduce bias while ensuring reliable training and evaluation conditions for anomaly detection and disease prediction models. The standardized feature vectors were

subsequently supplied to the Autoencoder and Random Forest models for latent feature extraction and classification. Normalization reduces feature variance and improves optimization efficiency, while engineered attributes capture hidden relationships among clinical variables. These preprocessing operations contribute significantly to stable model convergence, improved generalization capability, enhanced anomaly detection performance, and accurate disease prediction across heterogeneous healthcare datasets.

3.2.1 Data Cleaning and Normalization

Data cleaning procedures remove missing values, inconsistent records, duplicated observations, and noisy samples that may negatively influence predictive performance and model reliability. Healthcare datasets frequently contain incomplete patient information and measurement inconsistencies arising from clinical data acquisition processes. Therefore, preprocessing operations are essential for improving data integrity and reducing uncertainty during model training. Proper normalization improves optimization stability and prevents dominant variables from biasing model learning. In healthcare datasets, variables such as glucose concentration, insulin levels, tumor radius, and area measurements often exist on different numerical scales, making feature transformation essential for efficient learning and accurate anomaly detection.

The missing value replacement process is expressed as:

$$X_{imp} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (8)$$

where, Ximp = Imputed feature value Xi= Available observations of the feature N = Number of valid observations

Following cleaning and normalization, the datasets become more suitable for machine learning analysis and latent representation learning. The Autoencoder utilizes these standardized inputs to generate compressed feature representations while minimizing reconstruction error. Data normalization also accelerates convergence, improves optimization efficiency, and enhances the capability of anomaly detection models to

identify unusual healthcare patterns. Consequently, these preprocessing operations significantly improve classification accuracy, model robustness, and generalization performance across heterogeneous healthcare datasets.

3.2.2 Dataset Splitting and Preparation

The prepared healthcare datasets were partitioned into training and testing subsets for model construction, validation, and performance evaluation. Stratified sampling techniques were adopted to preserve the original class distribution across both subsets and ensure balanced representation of disease categories. Typically, 80% of observations were allocated for model training while the remaining 20% were reserved for testing purposes. This strategy enables unbiased estimation of model generalization capability and reduces the possibility of overfitting during optimization. Proper dataset preparation ensures that anomaly detection and classification algorithms are evaluated under realistic and representative healthcare environments containing diverse patient characteristics and disease conditions.

The training dataset size is represented as in eqn 9:

$$N_{train} = 0.8N \quad (9)$$

where, Ntrain = Number of training samples, N = Total number of observations

The testing dataset size is represented as in eqn 10:

$$N_{test} = 0.2N \quad (10)$$

where, Ntest = Number of testing samples, N = Total number of observations

After partitioning, the training subset is utilized for learning feature representations and optimizing model parameters, while the testing subset evaluates predictive performance on unseen data samples. Stratified sampling preserves disease prevalence ratios and prevents class imbalance from influencing evaluation metrics. This preparation process improves reliability of anomaly detection, supports fair comparison among predictive models, and ensures realistic assessment of disease classification performance in healthcare applications.

3.3 Autoencoder-Based Latent Feature Extraction

Autoencoder networks were employed to learn compact latent representations of healthcare records while preserving critical clinical information and reducing feature dimensionality. The Autoencoder consists of two major components: an encoder and a decoder. The encoder compresses high-dimensional healthcare features into a low-dimensional latent vector that captures the most informative patterns present in patient records, whereas the decoder reconstructs the original input from the compressed representation. This architecture

enables the model to identify hidden nonlinear relationships among clinical variables that may not be apparent in the original feature space. By learning meaningful representations, the Autoencoder improves feature quality and enhances anomaly detection capability in complex healthcare environments.
The encoding function is defined as in eqn 11:

$$Z = f(Wx + b) \quad (11)$$

where, z = Latent feature representation, W = Weight matrix, x = Input feature vector, b = Bias term, f(·)= Activation function
The encoder progressively transforms the original healthcare features into a compressed latent space representation by extracting dominant statistical and clinical patterns from patient records. This dimensionality reduction process eliminates redundant information while preserving discriminative characteristics necessary for disease prediction and anomaly identification. The learned latent vectors provide a compact summary of complex healthcare observations and significantly reduce computational complexity during subsequent learning stages. Moreover, the latent representation improves robustness against noise and missing values commonly encountered in real-world clinical datasets. Figure 2 illustrates the Autoencoder architecture used for latent feature extraction. The encoder compresses high-dimensional healthcare variables into informative latent vectors, while the decoder reconstructs original records from compressed representations. This process removes redundancy and noise while preserving clinically relevant information for downstream predictive tasks.

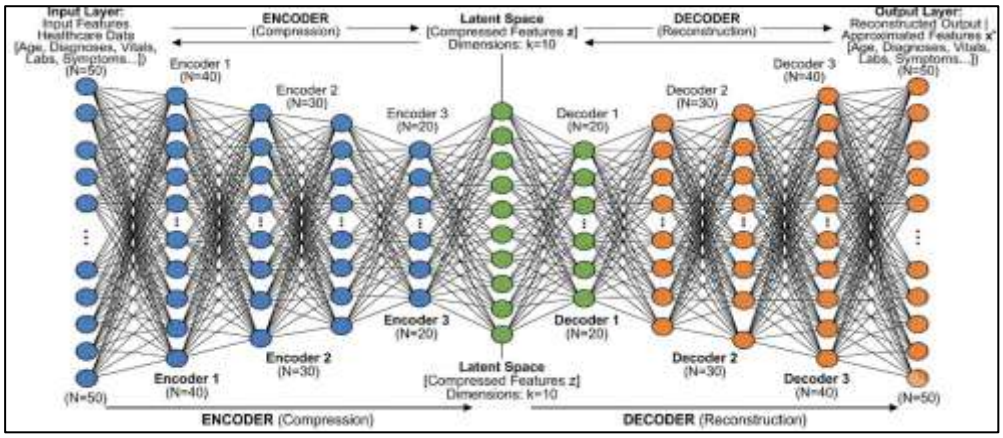


Figure 2: Autoencoder-Based Latent Feature Extraction Architecture for Healthcare Data

The reconstruction process is expressed as in eqn 12:

$$\hat{x} = g(W'z + b') \quad (12)$$

where, x̂ = Reconstructed feature vector, W′ = Decoder weights b ′ = Decoder bias g(·) = Decoder activation function
The Autoencoder training objective minimizes reconstruction loss to ensure accurate recovery of the original input features from the latent space. The compressed representations retain important clinical information while removing redundancy and noise from healthcare records. These latent features capture hidden nonlinear dependencies among patient attributes and provide discriminative information for downstream anomaly detection and disease classification tasks. Consequently, the extracted latent representations improve prediction accuracy, computational efficiency, and model generalization across heterogeneous healthcare datasets.

3.4 Hybrid Anomaly Detection and Classification Framework

The proposed framework integrates unsupervised anomaly detection with supervised classification to improve healthcare prediction reliability and diagnostic accuracy. Initially, the Autoencoder extracts low-dimensional latent representations from high-dimensional healthcare records while preserving important clinical characteristics. These latent features are subsequently supplied to anomaly detection algorithms, including Isolation Forest and K-Means clustering, to identify hidden anomalies and patient subgroups that may not be distinguishable using conventional classification methods. Isolation Forest detects abnormal healthcare observations based on feature isolation mechanisms, whereas K-Means clustering groups patients according to latent similarities in clinical characteristics. The outputs generated by these methods are combined with latent representations and supplied to the XGBoost classifier for final disease prediction and decision making.

The Isolation Forest anomaly score is represented as in eqn 13:

$$a = IF(z) \quad (13)$$

where, a = Isolation Forest anomaly score, IF(·)= Isolation Forest function, z = Autoencoder latent feature vector
The anomaly score generated by Isolation Forest quantifies the degree of abnormality associated with each healthcare observation in the latent feature space. Records with higher anomaly scores are considered more likely to represent unusual clinical conditions, rare disease manifestations, or potential data inconsistencies. Simultaneously, K-Means clustering organizes patient records into homogeneous groups based on similarities among latent features extracted by the Autoencoder. These cluster assignments reveal hidden patient subpopulations and support the identification of underlying healthcare patterns. The anomaly information and cluster memberships provide complementary knowledge that enhances disease prediction performance and model interpretability.
The final prediction function is represented as in eqn 14:

$$\hat{y} = F(z, a, c) \quad (14)$$

where, yˆ = Predicted disease label, z = Autoencoder latent features a = Isolation Forest anomaly score, c = K-Means cluster label, F(·)= XGBoost classification function
The hybrid learning framework combines deep feature extraction, anomaly detection, and supervised classification to improve predictive robustness and diagnostic reliability. The anomaly scores generated by

Isolation Forest assist in identifying unusual patient patterns, while K-Means cluster assignments capture hidden subgroup relationships within healthcare records. Integrating these complementary sources of information enables the XGBoost classifier to make more informed decisions and enhances disease prediction accuracy. Consequently, the proposed hybrid architecture improves generalization capability, reduces misclassification rates, and supports robust healthcare analytics in heterogeneous clinical environments.

3.4.1 Isolation Forest and K-Means Analysis

Isolation Forest identifies abnormal patient records by measuring the isolation depth of observations within randomly generated partition trees. Observations that require fewer partitions for isolation are considered anomalous because they differ significantly from the majority of healthcare records. In parallel, K-Means clustering groups similar patient records into clusters according to feature similarity in the latent representation space generated by the Autoencoder. The anomaly scores and cluster labels produced by these methods provide complementary information regarding hidden structures and relationships within healthcare data. Isolation Forest effectively identifies outliers and unusual disease manifestations, whereas K-Means captures population-level patterns and patient subgroup characteristics that are useful for predictive modeling and healthcare analytics.
The Isolation Forest anomaly score is computed as in eqn 15:

$$a_i = IF(z_i) \quad (15)$$

where, ai = Anomaly score of patient i, IF(·)= Isolation Forest function zi = Latent feature vector of patient i.
The Isolation Forest algorithm operates by recursively partitioning healthcare observations using randomly selected features and split values until individual samples become isolated. Since anomalous observations are relatively rare and differ considerably from normal patient records, they tend to be isolated with fewer partitioning operations and therefore receive higher anomaly scores. This characteristic makes Isolation Forest highly effective for detecting unusual disease manifestations, abnormal physiological measurements, and rare clinical conditions that may otherwise remain undetected in conventional supervised learning frameworks. The anomaly detection process is performed in the latent feature space generated by the Autoencoder, where complex nonlinear relationships among healthcare variables are already compressed into informative representations. Operating in this transformed space improves anomaly sensitivity and reduces the influence of noisy and redundant features. Furthermore, the anomaly scores generated by Isolation Forest provide valuable supplementary information regarding patient abnormality levels and risk profiles, which can subsequently be combined with clustering information and classification models to improve healthcare prediction reliability and diagnostic decision support performance. Figure 3 shows the anomaly characterization stage of the proposed framework. Isolation Forest identifies abnormal healthcare observations using random partition trees, whereas K-Means discovers patient subgroups according to latent feature similarity. Their complementary outputs improve anomaly awareness and support more robust healthcare predictions.

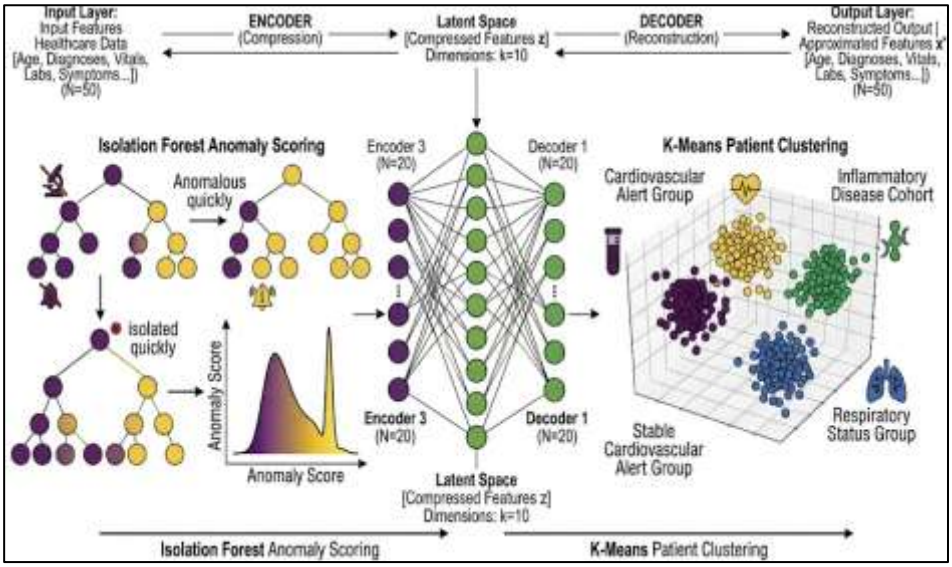


Figure 3: Hybrid Anomaly Characterization Using Isolation Forest and K-Means Clustering

The K-Means clustering assignment is represented as in eqn 16:

$$C_i = arg||z_i - \mu_k||^2 \quad (16)$$

where,

ci = Cluster label of patient I, zi = Latent feature vector μk = Centroid of cluster k, k = Number of clusters
The anomaly scores generated by Isolation Forest quantify deviations from normal healthcare patterns and assist in identifying rare clinical conditions. Meanwhile, K-Means clustering reveals hidden patient communities that share similar physiological or diagnostic characteristics. Combining these outputs with Autoencoder latent features improves anomaly characterization and enables the predictive model to utilize both local outlier information and global population structures. Consequently, the integrated framework enhances disease prediction accuracy, robustness, and interpretability across heterogeneous healthcare datasets.

3.4.2 LightGBM Based Classification

The proposed hybrid framework used LightGBM technique acts as eventual supervised classification. The classifier gets the latent representations resultant from the Autoencoder, then anomaly scores caused by the Isolation Forest, and cluster data attained over K-Means clustering. LightGBM achieves the best result by using a gradient boosting method to design many successive decision trees, with each following tree improving the errors built by the old trees. The structure is allowed to acquire the knowledge of non-linear association and combined involvement between the features. By which turn would it be convenient for disorder classification problems involving various electronic healthcare records.The LightGBM prediction function is represented as in eqn 17:

$$\hat{y} = \sum f_t(x), t = 1 \; to \; T \qquad (17)$$

where,ŷ = Predicted disease label ft(x) = Prediction of tree t, T = Total number of trees in the ensemble
The objective function minimized during training is expressed as:

$$L(t) = \sum l(y_i, \hat{y}(t-1) + f_t(x_i)) + \Omega(f_t), i = 1 \; to \; n \qquad (18)$$

where, L = Overall objective function, l(y_i,ŷ_i) = Prediction loss function
Ω(ft) = Regularization term for tree complexity, n = Number of training samples

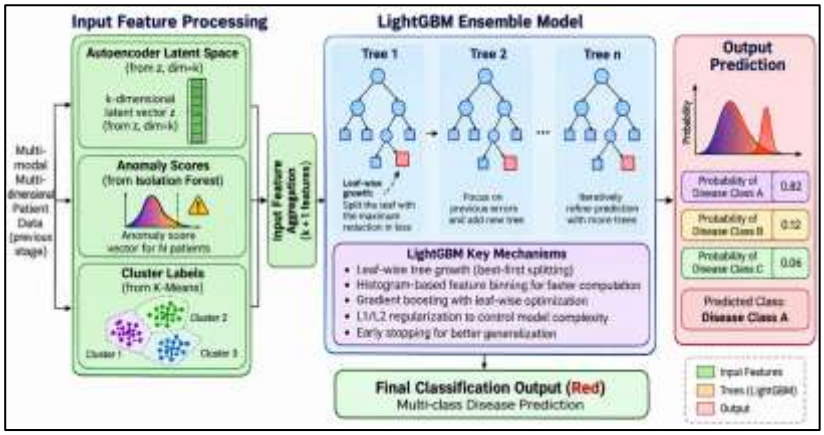


Figure 4: LightGBM Based Disease Classification Framework Using Latent and Anomaly Features

Figure 4 presents the LightGBM classification stage of the hybrid framework. The classifier merges latent features, anomaly scores, and cluster works to build the most optimal decision trees that strengthen disease discrimination, predictive robustness, and spread capability across diverse healthcare records..

4. Results And Analysis

In this study, multiple anomaly detection and hybrid machine learning models were evaluated using an expanded diabetics and cancer dataset containing approximately 2000 and 50,000 patient records. The primary objective was to investigate the effectiveness of standalone unsupervised anomaly detection approaches and hybrid frameworks integrating deep learning with ensemble classification techniques. Four major model configurations were considered, namely Autoencoder with Isolation Forest, Autoencoder with K-Means clustering, Autoencoder with Isolation Forest and XGBoost, and Autoencoder with K-Means and XGBoost. The Autoencoder was employed to generate compact latent feature representations from high-dimensional healthcare data, while Isolation Forest and K-Means provided anomaly characterization and patient subgroup identification. XGBoost was subsequently used as a supervised classifier to improve disease prediction performance. Model evaluation was conducted using Accuracy, Precision, Recall, and Area Under the Receiver Operating Characteristic Curve (AUC) metrics. These evaluation measures provide

comprehensive insights into predictive capability, class discrimination, and robustness under complex healthcare environments characterized by nonlinear feature relationships and heterogeneous patient populations.

4.1 Diabetes Dataset Results

The proposed hybrid anomaly-aware healthcare framework was first evaluated using the diabetes dataset containing 2,000 patient records and eight clinical attributes, including pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index, diabetes pedigree function, and age. The objective of this experiment was to investigate the effectiveness of standalone anomaly detection methods and Autoencoder-based hybrid models for diabetes prediction. Experimental results indicate that purely unsupervised approaches provide limited predictive capability because they operate without class labels and rely only on structural abnormalities present in the data. The AE + Isolation model achieved an accuracy of 52.0%, while AE + KMeans achieved 66.5% accuracy. Similarly, Isolation + KMeans produced nearly random classification performance with 50% accuracy and AUC. However, substantial improvements were observed after integrating XGBoost with Autoencoder-generated latent features. The AE + Isolation + XGBoost and AE + KMeans + XGBoost models achieved accuracies of 85.0% and 85.25%, respectively, with AUC values close to 0.93. These findings confirm that supervised boosting significantly improves anomaly-aware healthcare prediction performance.

4.2 Comparative Analysis with LightGBM

This table 2 shows Accuracy, Precision, Recall, F1-Score and ROC-AUC values for ten models on diabetes dataset with 2 000 records. Unsupervised models show close results to random: Autoencoder 53.50 % / 0.6231 AUC, Isolation Forest 52.75 % / 0.6036, KMeans 55.50 % / 0.5550, AE+Isolation 52.00 %, Isolation+KMeans 50.00 %. AE+KMeans achieves 66.50 %. Hybrid supervised models show tremendous growth: AE+KMeans+XGBoost 84.50 % / 0.9338 AUC, AE+Isolation+XGBoost 88.00 % / 0.9462. Two models based on LightGBM lead—AE+KMeans+LightGBM (proposed) 89.25 % accuracy, 0.8720 precision, 0.9200 recall, 0.8954 F1, 0.9492 AUC; AE+Isolation+LightGBM 90.75 % / 0.9593 AUC—the fact proves that latent features together with anomaly scores in LightGBM lead to best diabetes prediction.

Figure 5 provides grouped bar charts for Accuracy, Precision, Recall and F1-Score as well as separate ROC-AUC

bar chart for models on diabetes dataset. The left-hand grouped bar charts of unsupervised models are grouped between 50–67 % accuracy and AUC values of 0.50–0.66. Four hybrid models on the right hand side provide 84–91 % accuracy and 0.93–0.96 AUC values. AE+Isolation+LightGBM gives the highest bar charts (90.75 % accuracy, 0.9593 AUC), followed by the proposed AE+KMeans+LightGBM (89.25 %, 0.9492). The picture illustrates the performance leap provided by combining Autoencoder latent features with LightGBM model on the diabetes dataset.

Figure 6 shows Isolation Forest anomaly scores (green line) versus the number of sample (X axis) for approximately 400 records of diabetes data, red dots correspond to anomalies found. The anomaly scores oscillate from –0.12 to +0.10, anomalies are detected as the most negative peaks of the curve. Those points correspond to the patients' atypical clinical profiles that are isolated early in random tree of the unsupervised detector. The figure illustrates the use of unsupervised detector, applied to latent space of Autoencoder, that gives additional anomaly scores to be used in hybrid LightGBM classifier.

Table 2. Performance Comparison of All Models on the Diabetes Dataset

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Autoencoder	53.50	0.6750	0.1350	0.2250	0.6231
Isolation Forest	52.75	0.6341	0.1300	0.2158	0.6036
KMeans	55.50	0.6528	0.2350	0.3456	0.5550
AE+ Isolation	52.00	0.6300	0.1286	0.2053	0.5855
AE+KMeans	0.6650	0.7250	0.7038	0.7006	0.6650
Isolation + kmeans	0.5000	0.5304	0.1256	0.2105	0.5000
AE + KMeans + XGBoost	84.50	0.8194	0.8850	0.8510	0.9338
AE + KMeans + LightGBM	89.25	0.8720	0.9200	0.8954	0.9492
AE + Isolation + XGBoost	88.00	0.8551	0.9150	0.8841	0.9462
AE + Isolation + LightGBM	90.75	0.9055	0.9100	0.9077	0.9593

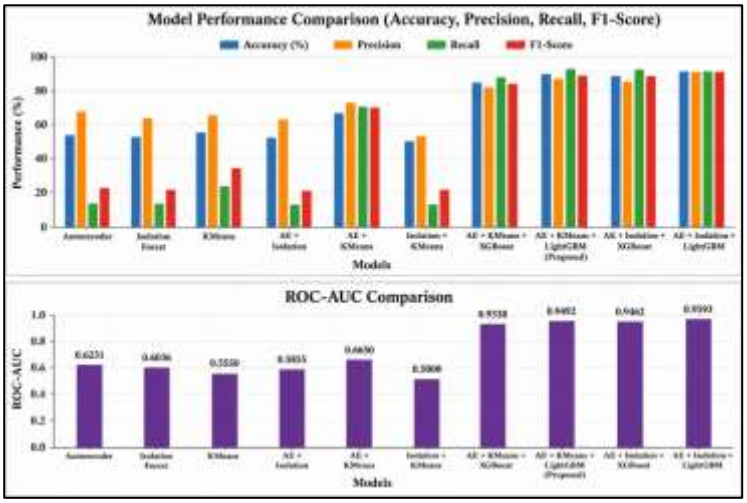


Figure 5: Performance comparison

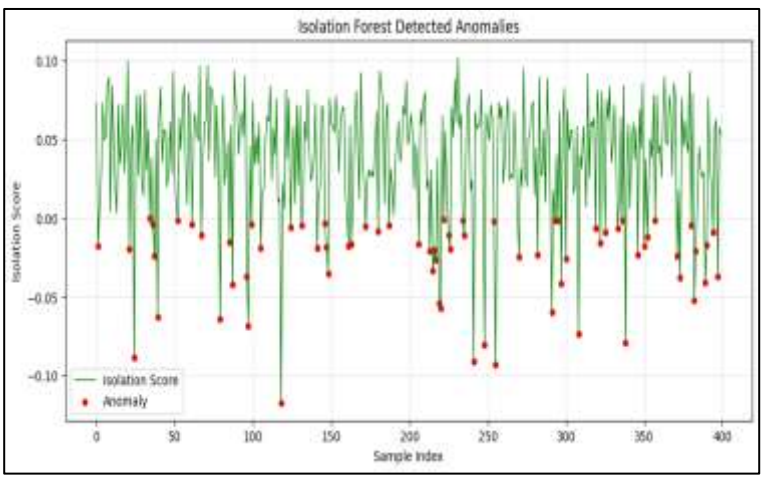


Figure 6: Isolation forest detected anomalies

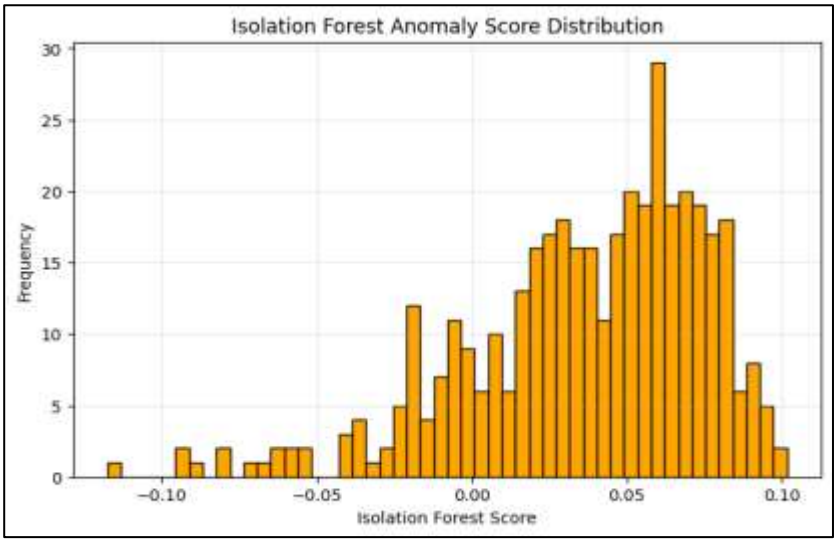


Figure 7: Isolation forest anomaly score

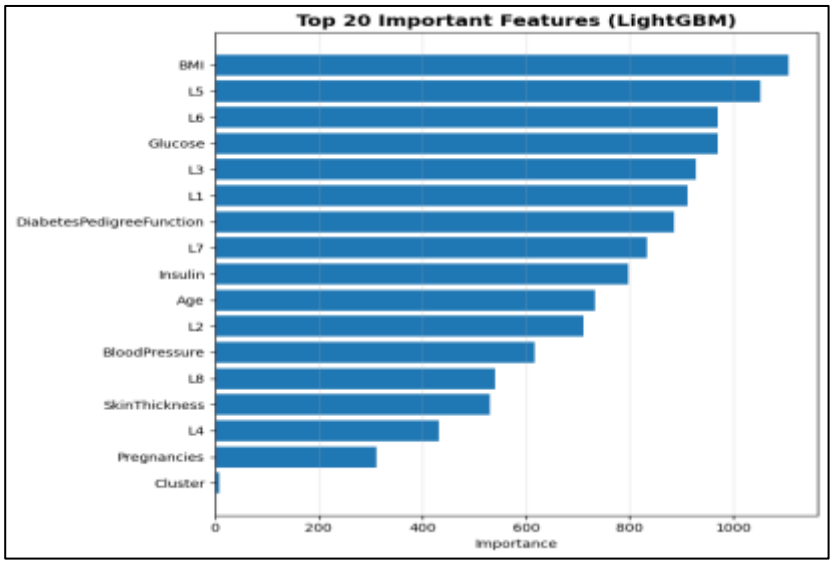


Figure 8: Features used in LightBGM

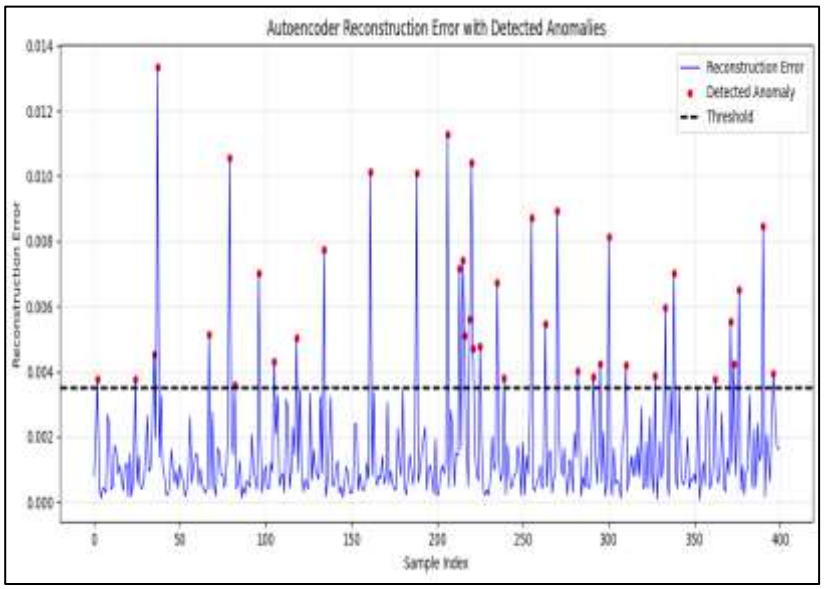


Figure 9: Autoencoder reconstruction error

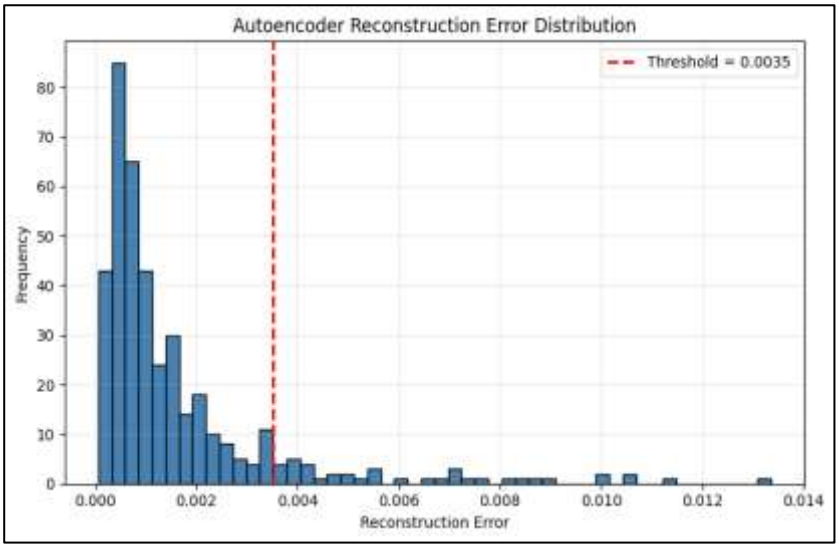


Figure 10: Error distribution of AE

Table 3: Performance Comparison of Standalone and Hybrid Models on the Cancer Dataset

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Autoencoder	49.97	0.4969	0.0531	0.0960	0.5511
Isolation Forest	52.01	0.5956	0.1248	0.2063	0.5712
KMeans	51.10	0.5550	0.1110	0.1851	0.5486
AE + Isolation	54.63	54.63	54.54	54.35	54.63
AE + KMeans	50.27	50.16	54.07	2.03	50.16
Isolation + KMeans	50.17	50.05	54.16	53.70	50.05
AE + KMeans + XGBoost	97.33	0.9683	0.9787	0.9735	0.9966
AE + KMeans + LightGBM	97.51	0.9705	0.9806	0.9759	0.9969
AE + Isolation + XGBoost	97.40	0.9700	0.9710	0.9611	0.9950
AE + Isolation + LightGBM	97.49	0.9698	0.9802	0.9750	0.9968

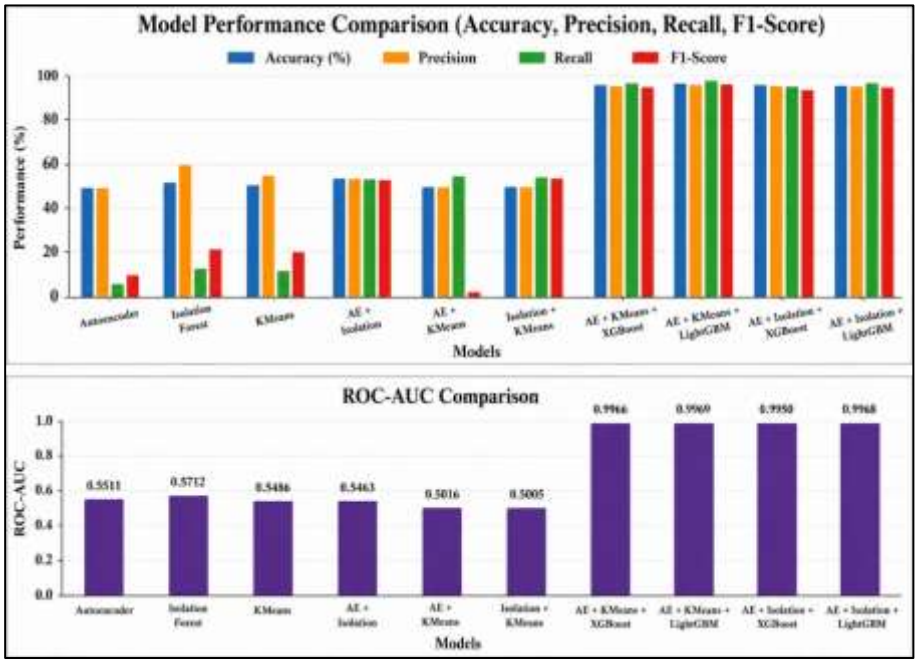


Figure 11: Comparison of Standalone and Hybrid Models on the Cancer Dataset

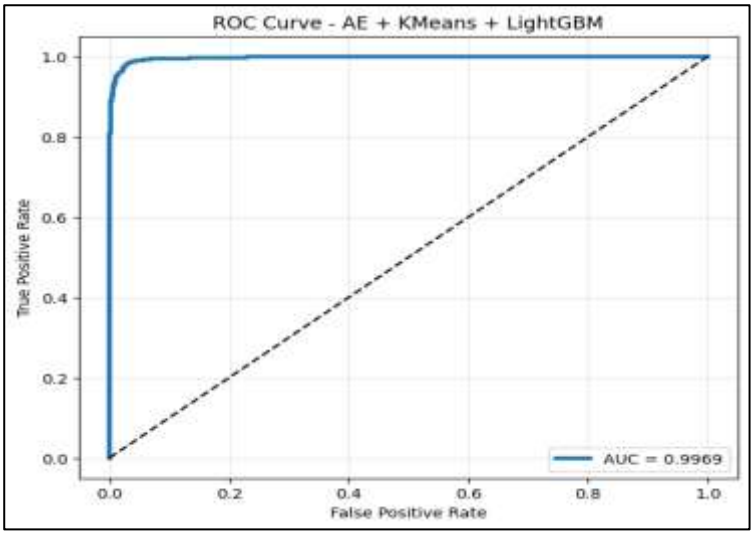


Figure 12: **ROC**

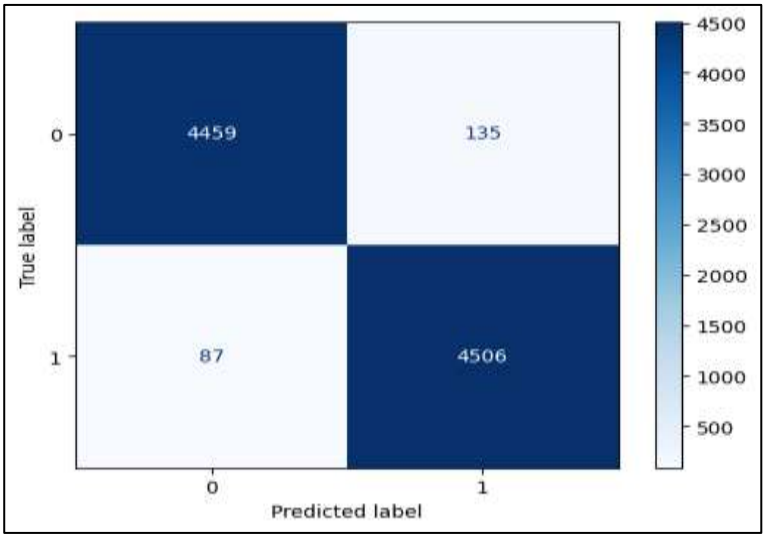


Figure 13: Confusion matrix

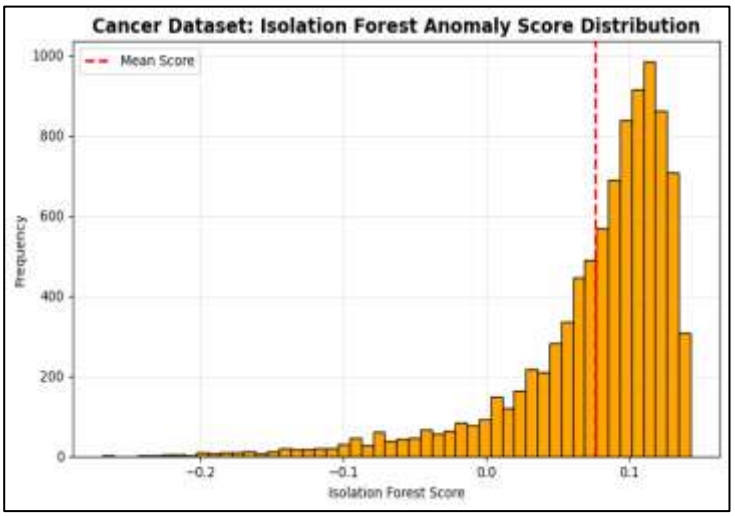


Figure 14: Cancer dataset isolation forest anomaly score

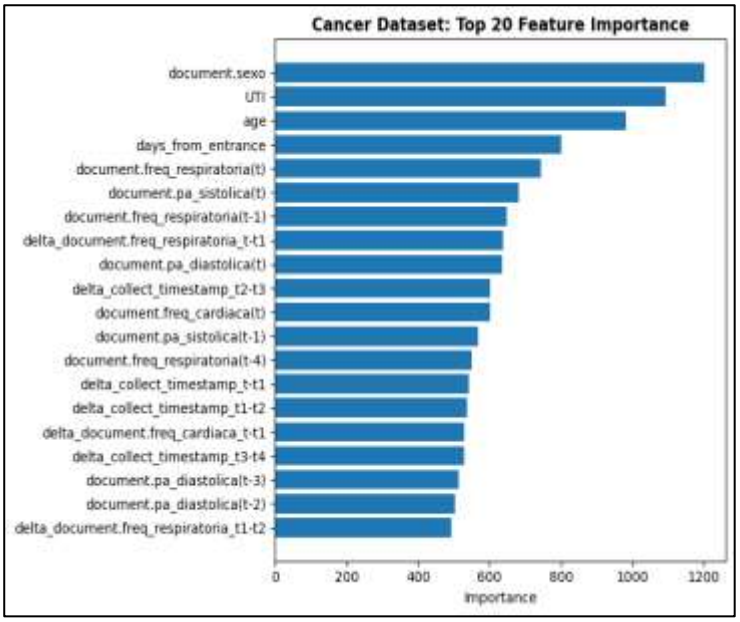


Figure 15: Cancer dataset top 20 features

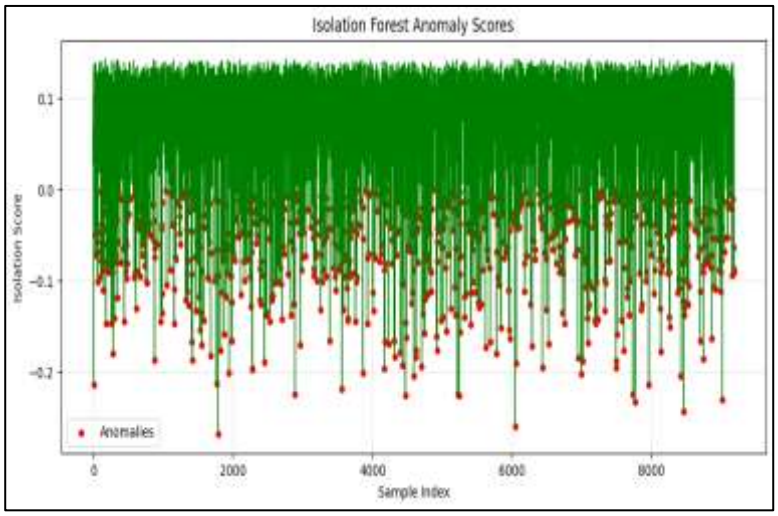


Figure 16:Isolation Anomaly Score

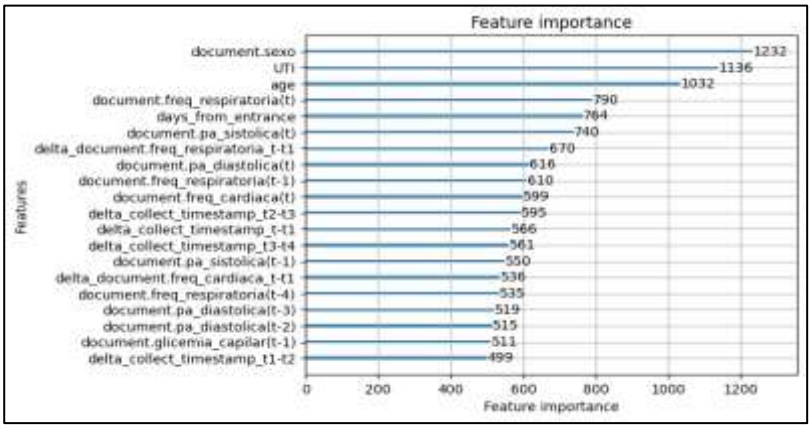


Figure 17: Feature importance

Figure 7 shows the histogram of Isolation Forest anomaly scores on the latent features of diabetes dataset. The histogram has unimodal shape with the centre at +0.05. It has a long left tail down to −0.12 and the peak of probability near +0.06 (the frequency of that peak ≈ 29). Majority of normal samples have positive anomaly scores, and the rest, which have negative scores, are anomalies. Those numerical anomaly scores serve as additional feature in the model of AE+Isolation+LightGBM, which leads to the highest ROC-AUC value (0.9593). Figure 8 shows top-20 features by importance for LightGBM on diabetes task. "BMI" is the leader (≈ 1 100),

followed by "L5", "L6" (≈ 1 050 and 980), "Glucose" (≈ 970), "L3", "L1", "DiabetesPedigreeFunction", "L7", "Insulin", "Age", "L2", "BloodPressure", "L8", "SkinThickness", "L4", "Pregnancies" and "Cluster" (near zero). Dominance of original features and latent dimensions of Autoencoder proves that hybrid model utilizes latent information together with classical risk factors.

Figure 9 shows the curve of reconstruction error of Autoencoder across ≈ 400 samples (blue line) with the horizontal threshold at 0.0035 (dashed black) and red dots corresponding to the samples with the reconstruction error above the threshold. Several spikes of reconstruction errors reach 0.013, and they are completely separated from the typical errors, lying in the band of 0.001–0.002. Those samples having high reconstruction errors are marked as anomalies and provide additional unsupervised signal to be used by the hybrid models.

Figure 10 shows the histogram of reconstruction errors of Autoencoder. Majority of samples form a narrow peak near 0.0005–0.001 (frequency > 80) with rapid decay; there is a threshold at 0.0035 shown by red dashed line. The histogram illustrates that Autoencoder reconstructs most normal patients with low error and marks anomalies by higher reconstruction error, thus providing unsupervised signal for the LightGBM classifiers.

Table 3 shows the same ten models on the 50 000-record cancer dataset. Unsupervised models demonstrate poor performance: Autoencoder 49.97 % / 0.5511 AUC, Isolation Forest 52.01 %, KMeans 51.10 %, AE+Isolation 54.63 %, AE+KMeans 50.27 %, Isolation+KMeans 50.17 %. Hybrid boosting models have near perfect results: AE+KMeans+XGBoost 97.33 % / 0.9966 AUC, AE+Isolation+XGBoost 97.40 % / 0.9950. Two

models based on LightGBM are nearly identical and are the best of all—AE+KMeans+LightGBM 97.51 % accuracy, 0.9705 precision, 0.9806 recall, 0.9759 F1, 0.9969 AUC; AE+Isolation+LightGBM 97.49 % / 0.9968 AUC.

Figure 11 is similar to Figure 5 for the cancer dataset: the grouped bar charts for Accuracy/Precision/Recall/F1-Score and the separate ROC-AUC chart. Unsupervised models are hovering around 50–55 % accuracy and AUC ≈ 0.50–0.57. Four hybrid models jump to 97 %+ accuracy and AUC values of 0.9950–0.9969. AE+KMeans+LightGBM shows the highest AUC (0.9969); AE+Isolation+LightGBM are virtually the same (0.9968). The picture illustrates dramatic advantage of using Autoencoder latent features together with LightGBM on the high-dimensional cancer dataset.

Figure 12 shows the ROC curve of AE+KMeans+LightGBM model on cancer test dataset. The curve rises almost vertically to 1.0 TPR at the FPR of nearly zero, then stays in the upper-left corner of the plot, which corresponds to the AUC value of 0.9969. Perfect separation of classes proves the outstanding discriminatory ability of the hybrid model when using latent features and clusters together with LightGBM.

Figure 13 shows the confusion matrix of the best hybrid model on the cancer dataset (approximately 9 187 samples). True negatives = 4 459, false positives = 135, false negatives = 87, true positives = 4 506. Extremely low off-diagonal counts confirm the 97.5 % accuracy, high precision/recall (> 0.97) and near-perfect F1-score values, which prove the clinical reliability of the AE+LightGBM pipeline.

Figure 14 shows the histogram of Isolation Forest anomaly scores on latent features of cancer dataset. The histogram has strong right-skew, with the peak near +0.12 (frequency ≈ 980) and long left tail down to –0.25. Red dashed line of mean anomaly score value is near +0.08. Minority of samples with negative anomaly scores are considered anomalies and they will further enrich the feature set of the AE+Isolation+LightGBM classifier.

Figure 15 shows the ranking of top-20 features used in LightGBM model on the cancer task. "document.sexo" (sex) is the leader (importance ≈ 1 232), followed by "UTI" (1 136), "age" (1 032), "days_from_entrance" (790), respiratory-frequency and blood-pressure measurements, and some other temporal delta features. Ranking illustrates the domination of demographic variables together with time-series physiological variables and Autoencoder latent information for the near-perfect classification results.

Figure 16 shows the Isolation Forest anomaly scores across the whole range of cancer dataset samples (≈ 0–9 000). Dense green band of normal anomaly scores lies between 0 and +0.15; red dots mark numerous extremely negative outliers (down to –0.25), which were detected as anomalies by the Isolation Forest. The visualization demonstrates the ability of the detector to find rare or atypical records that will be useful for hybrid LightGBM models.

Figure 17 shows another ranking of the top cancer features used by LightGBM. "Sex" is dominating (≈ 1 232), followed by "UTI", "age", "days from entrance", "respiratory frequency", "systolic pressure" and multiple lagged/delta vital-sign features. Consistency of the top ranks with Figure 20 proves that demographic and physiological time-series variables, together with Autoencoder latent information, are the key for achieving 97.5 % accuracy and 0.9969 AUC.

Discussion

Experimental results confirm that the Autoencoder is capable of identifying nonlinear correlations between attributes and providing compressed latent representations. Nevertheless, since these unsupervised approaches do not make use of the labels, their predictive performance is limited, and on both diabetes and cancer datasets, they generally provide accuracy values that range close to random or moderately improved classification performance.

On the other hand, the addition of supervised gradient-boosting classifiers allows achieving significant prediction improvements. XGBoost hybrid approaches yield promising results, where AE + K-Means + XGBoost achieves 84.50% accuracy and 0.9338 ROC-AUC on the diabetes dataset, while AE + Isolation + XGBoost achieves 88.00% accuracy and 0.9462 ROC-AUC. Yet, the LightGBM approaches achieve even better results, as AE + K-Means + LightGBM yields 89.25% accuracy and 0.9492 ROC-AUC, while AE + Isolation + LightGBM reaches the best diabetes results of 90.75% accuracy and 0.9593 ROC-AUC. A similar pattern in terms of performance was noted on the cancer dataset. In this case, AE + K-Means + XGBoost obtained an accuracy of 97.33% and ROC-AUC of 0.9966, while AE + Isolation + XGBoost received 97.40% accuracy and ROC-AUC of 0.9950. The configurations utilizing LightGBM provided slightly higher results with AE + K-Means + LightGBM reaching the best cancer accuracy of 97.51% and ROC-AUC of 0.9969 and AE + Isolation + LightGBM reaching 97.49% accuracy and 0.9968 ROC-AUC.

In general, the results indicate that LightGBM demonstrates the best predictive performance among all supervised boosting algorithms tested in this research, especially when coupled with latent representation of Autoencoder and anomaly or clustering information. Of all configurations suggested, AE + K-Means + LightGBM demonstrates the best performance in general and in consistent manner on both datasets with the highest cancer accuracy and ROC-AUC and at the same time highly competitive performance on diabetes dataset.

CONCLUSION

In this research work, the hybrid anomaly-aware prediction framework has been introduced by combining Autoencoder, Isolation Forest, K-Means clustering, and supervised gradient-boosting classifiers for the disease prediction and anomaly detection on diabetes and cancer data sets. This framework takes into account both latent feature extraction and anomalies within patients to enhance the reliability of health care predictions. From the experimental results, it is evident that the standalone unsupervised frameworks have less predictive capability since these frameworks only capture structural anomalies but not decision boundary of diseases.

The combination of supervised boosting technique enhances the predictive ability of the frameworks. The hybrid frameworks with XGBoost classifier perform much better than standalone anomaly detection techniques; however, the hybrid frameworks with LightGBM give best results. For diabetes data set, AE + Isolation + LightGBM framework gives highest accuracy of 90.75% and ROC-AUC of 0.9593. Similarly, for cancer data set, AE + K-Means + LightGBM gives highest accuracy of 97.51%, precision of 0.9705, recall of 0.9806, F1-score of 0.9759 and ROC-AUC of 0.9969. For this reason, it can be concluded from the experimental results that AE + K-Means + LightGBM is the most recommended configuration, offering the best performance among all the tested datasets, whereas AE + Isolation + LightGBM offers the best performance for the diabetes dataset. The findings clearly indicate that the combination of Autoencoder latent space along with anomalies or clusters and LightGBM supervised classification can significantly enhance disease prediction performance. The proposed approach opens up new possibilities for future healthcare analytics solutions in terms of anomaly detection on cloud database, privacy-preserving learning, federated healthcare, and security-based clinical intelligence.

Reference

1) Aggarwal, C. C. (2016). Outlier ensembles. In Outlier Analysis (pp. 185-218). Cham: Springer International Publishing.
2) Kalange, O. S., Kahat, R. S., Kale, A. S., Kale, T. R., & Joglekar, P. S. (2022). Implementation of Various Machine Learning Algorithms for Traffic Sign Detection and Recognition.
3) Fouladvand, S., Noshad, M., Goldstein, M. K., Periyakoil, V. J., & Chen, J. H. (2023). Mild cognitive impairment: data-driven prediction, risk factors, and workup. AMIA Summits on Translational Science Proceedings, 2023, 167.
4) Chollet, F., & Chollet, F. (2021). Deep learning with Python. simon and schuster.
5) Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.
6) Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.
7) Jolliffe, I. (2025). Principal Component Analysis. In International Encyclopedia of Statistical Science (pp. 1945-1948). Berlin, Heidelberg: Springer Berlin Heidelberg.
8) Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
9) MacQueen, J. (1967). Multivariate observations. In Proceedings ofthe 5th Berkeley symposium on mathematical statisticsand probability (Vol. 1, pp. 281-297). Oakland, CA, USA: University of California press.
10) Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
11) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
12) Samariya, D., Ma, J., Aryal, S., & Zhao, X. (2023). Detection and explanation of anomalies in healthcare

data. Health Information Science and Systems, 11(1), 20.
13) Utkin, L., Ageev, A., Konstantinov, A., & Muliukha, V. (2022). Improved anomaly detection by using the attention-based isolation forest. Algorithms, 16(1), 19.
14) Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. IEEE transactions on knowledge and data engineering, 35(12), 12591-12604.
15) Yadav, K., Aswal, U. S., Saravanan, V., Dwivedi, S. P., Shalini, N., & Kumar, N. (2023, December). Isolation forest anomaly detection in vital sign monitoring for healthcare. In 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI) (Vol. 1, pp. 1-7). IEEE.
16) Zhang, C., & Ma, Y. (2012). Ensemble machine learning (Vol. 144). New York: springer.
17) Zhou, Z. H. (2021). Machine learning. Springer nature.
18) Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. Nature genetics, 51(1), 12-18.
19) Najjar, R. (2023). Redefining radiology: a review of artificial intelligence integration in medical imaging. Diagnostics, 13(17), 2760.
20) Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., ... & Pierrot, T. (2025). Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 22(2), 287-297.
21) Preethi, P., & Asokan, R. (2019). An attempt to design improved and fool proof safe distribution of personal healthcare records for cloud computing. Mobile Networks and Applications, 24(6), 1755-1762.
22) Bharathy, S. S. P. D., Preethi, P., Karthick, K., & Sangeetha, S. (2017). Hand gesture recognition for physical impairment peoples. SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE), 610.
23) Asokan, R., & Preethi, P. (2021). Deep learning with conceptual view in meta data for content categorization. In Deep Learning Applications and Intelligent Decision Making in Engineering (pp. 176-191). IGI Global Scientific Publishing.
24) Sujithra, M., Velvadivu, P., Rathika, J., Priyadharshini, R., & Preethi, P. (2022, October). A study on psychological stress of working women in educational institution using machine learning. In 2022 13th international conference on computing communication and networking technologies (ICCCNT) (pp. 1-7). IEEE.
25) Kala, R., & Savitha, K. K. (2025, January). Anomaly Detection in Healthcare: Deploying Hybrid Machine Learning Techniques. In International Conference on Mathematical Modeling and Computational Science (pp. 467-479). Cham: Springer Nature Switzerland.