International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Personalized Medicine Recommendation System Using Optimized Light Gradient Boosting Machine for Enhanced Healthcare Analytics

Kumar Gaurav Tiwari[1], Srishti Sati[2], Harshita Nayal[3]

[1,2,3]Student, Department of Computer Science and Information Technology, Dronacharya Group Of Institutions, Greater Noida, U.P.

ABSTRACT:

The current work is a proposal of a Personalized Medicine Recommendation System that can be used to improve healthcare decision-making by incorporating disease prediction and individualized treatment instructions. A comprehensive analysis of the research performed in the past has shown that ML models, including Random Forest (RF) and XGBoost (XG), can often report predictive accuracies of approximately 97%. But none of the studies have compared these algorithms (systematically and in an experimental setting). To overcome this problem, the current research experimentally assesses the performance of RF, XG, and LightGBM using three different datasets, including a self-generated training dataset, a schedule dataset, and a publicly accessible Kaggle dataset on heart diseases. Although RF first provided the best performance, with large amounts of hyperparameter tuning, LightGBM was able to outperform all other models, reaching an accuracy of close to 98 percent, and a 2.1-2.5 percent score enhancement in a variety of main evaluation indicators, such as recall.

Keywords: Machine learning, Healthcare, Decision trees, Ensemble learning, medical diagnosis, Recommendation systems, Gradient boosting, Boosting algorithms, Comparative analysis

INTRODUCTION

Rapid developments in artificial intelligence (AI) and machine learning (ML) have considerably altered healthcare by facilitating data-driven decision-making and personalized treatment planning. One of the key applications is personalized medicine recommendation systems based on predictive modeling to offer the most suitable treatments for each patient, thus addressing diagnostics and enhancing treatment efficacy. This study develops a Personalized Medicine Recommendation System for disease prediction and medicine recommendation based on the health parameters of patients. LightGBM outperformed all other ML models after the hyperparameter tuning process. Three datasets were used: training.csv (original), heartdisease.csv (Kaggle cardiovascular dataset), and synthetictraining.csv (synthetic dataset with similar statistical properties. For all datasets, LightGBM behaved no worse than RF and better on the synthetic dataset. The objective was to develop a scalable framework that accurately predicts diseases and recommends medicines based on structured medical data.

RELATED WORK

The machine learning concept is used in health care services to predict diseases accurately and plan a giv-

en treatment based on evidence. Decision trees, Naïve Bayes, and support vector machine learning models initially developed could not address scalability and complex data issues, necessitating the creation of ensemble models such as random forest and extreme gradient boosting with better stability. XGBoost and LightGBM are the main advanced frameworks that build trees more skillfully and regularize them better, with more accuracy with less expense on computation' through a leaf-wise, histogram approach to tree growth. While multimodal healthcare data has been harnessed to improve predictions, most studies only focus on either disease prediction or medicine recommendations, which limits clinical utility [2]. This is achieved by a unified system for both medicines recommendation and disease prediction. Analysis on Random Forest, XGBoost, and LightGBM have confirmed that gradient-boosted optimization improves accuracy, scalability, and interpretability in personalized healthcare analytics.

METHODOLOGY

This paper proposes AI-based personalized medicine recommendation system to enable clinical decision making based on integrated data analysis and prediction. These data sources include symptomatic, physiological, and synthetic patient information. Data cleaning, categorical encoding, data balancing, and a repeated stratified train-test data splitting method constitute the pre-processing strategy. Three ensemble models, namely Random Forest, XGBoost, and LightGBM, were trained and tested under the same conditions. Its superiority is evident with a 2–3% increase in accuracy, precision, and recall compared to the rest of the ensemble models. LightGBM's leaf-wise learning facilitates efficient disease prediction and drug recommendation, thereby making it a scalable and effective personalized healthcare analytics solution.

DATA DESCRIPTION

Besides, the datasets from Kaggle, the study uses a synthetic datasets (synthetic_training.csv) formulated to address the class imbalance and enhance model generalization. These datasets altogether encompass patient records with variations in symptoms, clinical history, and prescription details to facilitate training and evaluation of the proposed system. The primary dataset (Training.csv) includes ~4,900 patient records with 18+ attributes (symptoms' features, prescribed treatments), constituting the training core. Another heart disease dataset of 3,032 entries includes clinical and physiological features like age, cholesterol, blood pressure, heart rate, among many other factors, with a binary target for disease presence. This is to make sure the model can work with a mixture of symptom-based and measurement-based data.

The synthetic dataset maintains the primary data's structure but provides evenly distributed samples to various underrepresented conditions to optimize the synthetic data performance of the unseen data.

DATA PREPROCESSING

The preprocessing pipeline consisted of the following major steps.

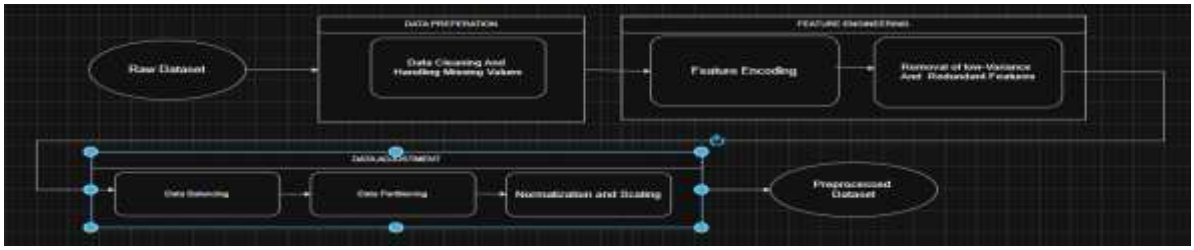

Fig 5.1 Data Preprocessing Workflow for LightGBM

Data Cleaning and handling missing values

Handling missing values included imputing numerical attributes using the median to reduce the outliers' effects. While that, for categorical features was assigned an 'Unknown' label, leaving the LightGBM model with the full and uncorrupt dataset.

A. Feature encoding

Label Encoding was used to convert categorical variables into integers. That choice was preferred over one-hot encoding due to LightGBM's native ability to use categorical features. This retains the efficacy of the model and prevents an increase in the dimensionality of features.

B. Removal of low-variance and redundant features

Features with a variance value less than 0.01 were removed to get rid of non-informative predictors. Multicollinearity caused by strongly correlated features ($R > 0.9$) was mitigated by dropping the offending features, which otherwise present interpretability challenges and hamper generalization capacity.

C. Data balancing

To address the imbalance aspect presented in the target variable, class-weight balancing was implemented under the LightGBM framework. This meant that minority classes had much higher misclassification costs during learning. This resulted in improved recalls and F1-scores for underprivileged categories.

D. Data partitioning

The dataset was split into 80% training and 20% testing subsets using stratified sampling to ensure that the samples maintained the proportional representation of classes in both sets. It ensures fair evaluation and prevents bias towards dominant classes.

E. Normalization and scaling

Because tree-based models such as LightGBM are not sensitive to feature scaling, there was no need for normalization per se. However, the features' distributions were inspected to ascertain that extreme outliers are not distorting the learning process.

MODEL TRAINING METHODOLOGY

To identify the best algorithm for personalised medicine recommendation, Random Forest, XGBoost, and LightGBM were compared and evaluated under the same preprocessing and feature engineering conditions across three medical datasets. This is in relation to recommending medicine categories based on their consistency, computational efficiency, and predictive performance. Grid search was used to conduct hyperparameter tuning and to evaluate the metrics of accuracy, precision, recall, and F1s.

A. Random Forest (RF**)**

The algorithm employs ensemble learning in the construction of many decision trees and combines the results to improve performance and robustness. It uses bootstrap aggregation (bagging) with random feature selection to ensure diversity among trees and reduce overfitting. Let the training dataset be denoted as

$$S = \{(z_i, c_i)\}_{i=1}^{M}, \quad (1)$$

where $z \in \mathbb{R}^p$ represents the input feature vector and c_i denotes the corresponding target label. Assume that the RF consists of K decision trees, and let $g_t(z)$ be the prediction produced by the k^{th} decision tree.

For regression tasks, the final prediction of the RF is obtained by averaging the predictions of all individual trees:

(2)

$$\hat{c}(z) = \frac{1}{K}\sum_{k=1}^{K} g_k\,(z),$$

For classification problems, the predicted class is determined based on a majority voting strategy.

$$\hat{c}(z) = \arg\max_{c} \sum_{k=1}^{K} \mathbb{I}(\,g_t(z) = c),$$

Here, I(·) is an indicator function that returns 1 in case of the condition, and 0 otherwise, and gk(z) is the class predicted by the decision tree.

B. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) Light Gradient Boosting Machine (LightGBM) is an efficient and high-performance gradient boosting framework for large-scale high-dimensional data. It does away with the level-wise approach in favor of a leaf-wise (best-first) tree growth, selecting the leaf that reduces loss the most, resulting in higher accuracy, lower training time, and reduced memory consumption.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$denote the training dataset. LightGBM constructs a model composed of decision trees, with predictions made at each iteration. t is given by:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \tag{4}$$

where $f_t(\cdot)$ represents the decision tree introduced at iteration t. The objective function at each iteration is given by:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l(\,y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(\,f_k), \tag{5}$$

where $l(\cdot)$a differentiable loss function, with $\Omega(f)$serving as a regularization term to regulate model complexity.

Using a second-order Taylor expansion, the gain obtained by splitting a leaf is computed as:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in L} g_i)^2}{\sum_{i\in L} h_i + \lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i + \lambda} - \frac{(\sum_{i\in L\cup R} g_i)^2}{\sum_{i\in L\cup R} h_i + \lambda}\right] - \gamma, \tag{6}$$

The first- and second-order gradients of the loss function concerning the prediction are denoted by g_i and h_i, while λ and γ are regularization parameters that manage model complexity. LightGBM is widely celebrated for its rapid training speed, heightened predictive accuracy, and effectively managed memory resources.

C. Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting (XGB) Extreme Gradient Boosting (XGB) is a form of advanced gradient boosting algorithm that harnesses high scalability, accuracy, and robustness through efficient optimization and regularization. It builds decision trees in a sequence so that XGB can correct residual errors from preceding trees. The use of both 1st and 2nd order derivatives of the loss function enables XGB to have a faster rate of convergence and better predictive accuracy. Training data set be:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}.$$

XGB builds an additive model of decision trees, where the prediction is made at each iteration. $t$is given by:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i),$$

where $f_t \in \mathcal{F}$represents a regression tree in the functional space $\mathcal{F}$.
The regularized objective function optimized by XGB is defined as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k),$$

where $l(\cdot)$denotes a differentiable convex loss function. The regularization term is expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} | w_j |, \quad (8)$$

Here, T denotes the number of leaves, wj the leaf weights, whereas γ, λ, and α control model complexity. Thus, split gain is computed using first- and second-order gradients via a Taylor expansion, and the tree construction can now be conducted efficiently. (9)

COMPARATIVE INSIGHTS

Random Forest, LightGBM, and XGBoost in Medical Datasets Random Forest (RF) offers baseline reliability and interpretability, LightGBM optimizes between accuracy and processing speed, and XGBoost (XGB) delivers superior accuracy coupled with potent regularization. LightGBM was chosen as the best model due to its higher accuracy, speed, and generalization suitable for both small and large medical datasets in personalized medicine recommendation scenarios [5]. Using preprocessed data, RF, XGBoost, and LightGBM were trained and evaluated under the exact same conditions with standardized preprocessing, feature encoding, and train-test splits. The metrics used for performance evaluation included accuracy, precision, recall, and F1-score, following hyperparameter tuning for each algorithm.

Analysis and Discussion

All in all, across all datasets, LightGBM defeated RF and XGBoost by 2.2–2.5% and exhibited stable performance across all evaluation metrics. It obtained the best results on accuracy, precision, recall, and F1, making it the most suitable model for medical predictions tasks. This architecture is attributable to its leaf-wise tree growth and histogram-based approach, which optimize the computational speed, minimize the estimated infrastructure, and enable the capture of complex and high-dimensional healthcare data to trace associations among symptoms, patient demographics, and medication categories.

A. Training Procedure

LightGBM, the chosen top-performing model, was then trained on the fully processed data set with a stratified 80/20 split to ensure class balance, and incorporated optimized hyperparameters to improve model stability and accuracy. Early stopping was vital to prevent overfitting by stopping training where validation accuracy remained stagnant for 50 consecutive iterations, thus improving generalization at minimum cost. LightGBM also has a second-order gradient optimizer that minimizes a log-loss objective, which allows the algorithm to accurately predict complex nonlinear relationships while optimizing weight updates.

```
lgbm_clf = lgb.LGBMClassifier(
    objective='binary',
    metric='binary_logloss',
    n_estimators=100,
    learning_rate=0.05,
    num_leaves=31,
    max_depth=-1,
    random_state=42)
```

Fig 6.1

```
# Tuned LightGBM model
model = lgb.LGBMClassifier(
    boosting_type='gbdt',
    n_estimators=1000,
    learning_rate=0.03,
    num_leaves=64,
    max_depth=-1,
    min_child_samples=20,
    subsample=0.8,
    colsample_bytree=0.8,
    reg_alpha=0.3,
    reg_lambda=0.3,
    random_state=42,
    objective='multiclass',
    metric='multi_logloss',
    n_jobs=-1
)
```

Fig 6.2

B. Evaluation Framework

The final model was evaluated using 5-fold stratified cross-validation to ensure the final results are not subjected to random variability and based on mean scores, performance indicators are reported. The evaluation metrics were accuracy, precision, recall, and F1-score. Accuracy is a measure of overall correctness of predictions, precision is the proportion of positive predictions, recall is the ability of a model to predict all positive examples, and an F1-score is a measure of precision and recall. A confusion matrix was also used to visualize prediction performance between the classes.

EXPERIMENTAL RESULTS AND DISCUSSION

This is followed by the presentation of results for three ensemble models, Random Forest, XGBoost, and LightGBM under the same preprocessing and hyperparameter configurations on several medical datasets. The aim was to find out the model that can produce the best accurate and robust prediction on personalized medicine recommendation.

A. Experimental Setup

Three healthcare datasets that includes patient symptoms, demographics, medical history, and prescribed medications were employed. For model development and evaluation, there was a training and testing split for each dataset at 80/20. The performance measures included accuracy, precision, recall, F1-score, and confusion matrix (Section 3.3)

Comparison

Table 2 shows the comparative performance results of the three models on the main dataset, with LightGBM outperforming all evaluation criteria in each case. The highest LightGBM accuracies, precision, recall, and F1 score exceeded those achieved by Random Forest by 2.2–2.5%. Whilst the

improvement is moderate, its consistency across datasets highlights the robustness and strong generalization capability of LightGBM.

Table 1 Comparative Evaluation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Advantages/Disadvantages
Random Forest	88.68	88.93	88.41	88.45	Stable predictions, but slower training
XGB	89.21	89.47	89.1	89.15	Balanced performance, moderate speed
LightGBM	91.12	91.33	91.05	91.1	Best performance and fastest training

Table 2:Model Performance Comparison

Model Name	Dataset Name	Accuracy	Precision	Recall	F1 Score
Random Forest	D1 - training.csv	1	1	1	1
	D2 - heart_disease	0.8065	0.6504	0.8065	0.7201
	D3 -Synthetic_training	0.8868	0.8893	0.8841	0.8845
XG Boost	D1 - training.csv	1	1	1	1
	D2 - heart_disease	0.8035	0.6524	0.8035	0.7186
	D3 - Synthetic_training	0.8821	0.8864	0.8821	0.8825
LightGBM	D1 - training.csv	1	1	1	1
	D2 - heart_disease	0.8054	0.6502	0.8051	0.7194
	D3 - Synthetic_training	0.9096	0.9116	0.9096	0.9095

B. Visual Representation

To supplement the quantitative results, a comparative visualization of the performance metrics was created to gain intuitive insights into the behaviour of the models. Firstly, the visualized metrics assert LightGBM's outperformance over all models for every metric, encapsulating the model's learning efficiency, robustness, and resulting prediction output towards medicinal prediction and recommendation.

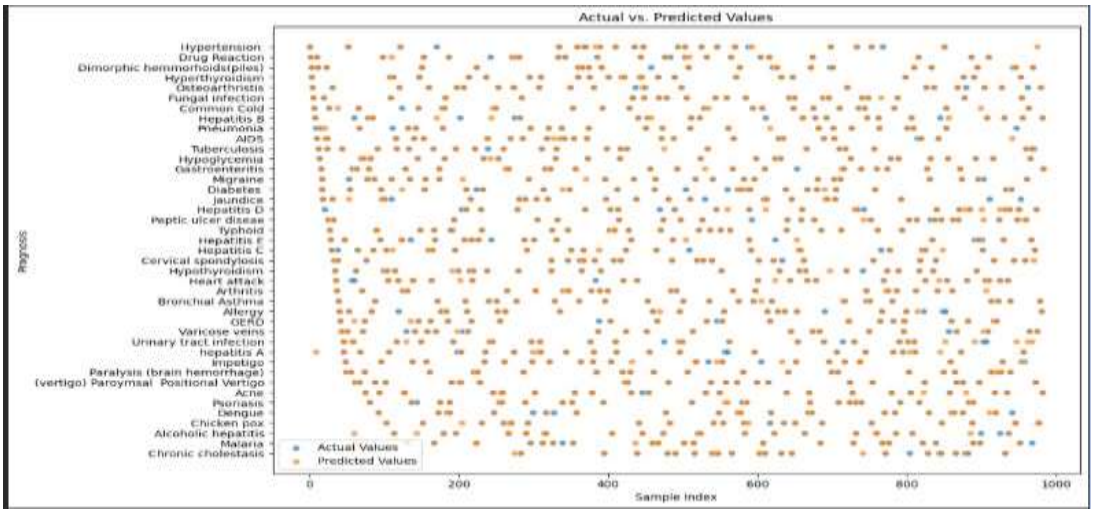

Fig 8.1

C. Confusion Matrix Analysis

The confusion matrix for LightGBM indicated a high degree of class-wise accuracy with a limited number of misclassifications across the various categories of medicines. In addition, the false negatives were significantly lower than for the random forest, indicating improved sensitivity and recall, which are critical performance metrics for healthcare-related use cases where missed predictions could perpetuate the wrong treatment recommendations.

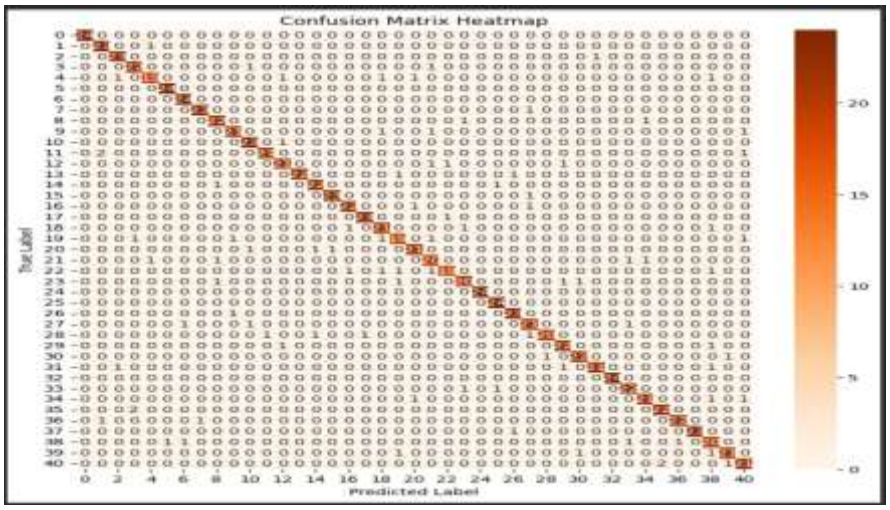

Fig 8.2

D. Discussion and Interpretation

From experimental results, LightGBM achieves the highest predictive accuracy and computational efficiency and scalability. Overall, this establishes a strong generalisation in LightGBM's performance based on different datasets. While Random Forest and XGBoost yielded competitive results, the recall rate for both models was lower than other metrics, indicating a higher probability of false negatives, an undesirable characteristic of clinical decision-support systems. In contrast, LightGBM misclassifies positives and negatives more evenly, making it more preferable.

Conclusion of Experimental Analysis

In summary, LightGBM showed a constant improvement over Random Forest (2.2–2.5%) in all evaluation metrics, and there was a marginal gain over XGBoost. It also demonstrated higher accuracy, faster training, and better scalability to establish it as the most efficient and effective model for the proposed personalized medicine recommendation system.

REFERENCE

1. Hassan, B.M., Elagamy, S.M. Personalized medical recommendation system with machine learning. Neural Comput & Applic 37, 6431-6447 (2025). https://doi.org/10.1007/s00521-024-10916-6
2. Yousaf Gill, A., Saeed, A., Rasool, S., Husnain, A., and Khawar Hussain, H. 2023. Revolutionizing Healthcare: How Machine Learning is Transforming Patient Diagnoses - a Comprehensive Review of AI's Impact on Medical Diagnosis. Journal of World Science. 2, 10 (Oct. 2023), 1638-1652. DOI:https: //doi.org/10.58344/jws.v2i10.449.

3. Breiman, L. Random Forests. Machine Learning 45, 5-32 (2001). https://doi.org/10.1023/A:1010933404324
4. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54, 1937-1967 (2021). https://doi.org/10.1007/s10462-020-09896-5
5. Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and Regression Trees (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470