Research Article

Benchmarking LSTM, Lightgbm, And Xgboost For Enhanced Cardiac Arrest Predictive Performance

Ms. R Padma Sahithi[1] , Ms. B Yamini Supriya[2]

[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, email: sahithiramayanam@gmail.com
[2]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, email: yamini.bommisetti@gmail.com

Abstract

Since cardiac arrest is a fatal medical emergency with a high death rate, it's critical to find efficient ways to anticipate it and take action as soon as possible. Despite the shortcomings of conventional risk assessment methods, ML is a rapidly emerging technology that promises to analyse complex clinical data to identify patients who are at risk.

To address this problem, we choose and compare three state-of-the-art ML models in this paper: LSTM, LightGBM, and XGBoost. These will be the main focus for predicting the cardiac arrest episodes with structured patient data.

During this process both models are trained and tested on the whole data set, and are assessed based on five metrics: accuracy, precision, recall, F1 score, and AUROC. In our comparison, we were able to find out what each model offers. The overall accuracy and AUROC scores were best for LightGBM. In a clinical environment, it is important for the algorithm to have the lowest false positive rate among them, which in this case was XGBoost. LSTM's highest recall score, on the other hand, shows that it identifies the most authentic positive cases. Results indicate that none of the models is superior for all. Instead, to make clinical decision support systems more reliable, a multi-model approach such as the ensemble technique could take advantage of the different strengths of these models. This study's findings add to the body of research and offers valuable data for future research and application in the healthcare field.

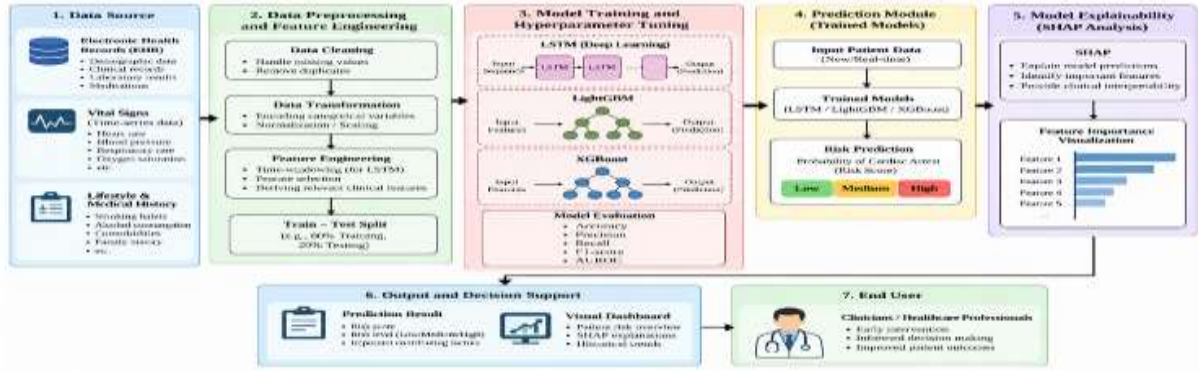
"Keywords— Cardiac Arrest Prediction, machine learning, Gradient Boosting, LSTM, Benchmarking."

I. INTRODUCTION

Sudden death, that is, without warning, and which represents a substantial proportion of the disease burden to health systems, remains a large cause of deaths worldwide. Still, early detection remains difficult due to the complex linkages of clinical data which are not always reflected in the traditional risk assessment tools [1, 2] Early recognition of cardiac arrest can give these facilities the maximum benefit with appropriate hospital resource utilization, rapid initiation of appropriate therapeutic interventions, and improve the chances of survival [4, 6].

This demand for more complex modelling has resulted in a significant trend towards machine learning approaches. Lately, ML has been investigated as a promising method which can detect early warning signs, learn complex non-linear patterns from large heterogeneous data sets, static clinical data, and dynamic physiological data [3, 5]. By incorporating demographic information, comorbidities, and lifestyle factors to create individualized risk assessments, ML models in cardiovascular health have also demonstrated the potential to go beyond conventional risk scores [9], [8].

The careful design of the experimental procedure made it possible to make a direct comparison between the different modeling paradigms. The Fig. summarizes the entire process, from feature engineering to review. First. Then, the three competing algorithms - LightGBM, XGBoost and the LSTM DL algorithm - are trained and fine-tuned under the same controlled conditions and the best algorithm is used for a final SHAP analysis for better clinical interpretation of predictions.



"Fig.1: Research Methodology Workflow and Comparative Modeling Pipeline"

1.1 Deep Learning Approaches for Time-Series Analysis

One of the important topics is sequence dependency, for instance LSTM network [16]. Recently LSTMs have been shown to work well in processing high-frequency data streams like EHR event sequences or continuous monitoring of vital signs, aiming to detect subtle changes in time leading up to a major event like cardiac arrest [19]. For instance, previous work demonstrated the ability to predict cardiac decompensation in the general ward in real time using DL models, and to learn patient decompensation trends over time [19, 4]. When the sensitivity of the prediction (Recall) is more crucial, the LSTM is a good choice, as it is designed to gain the more complicated and multi-layered data [5].

1.2 The Efficacy of Gradient Boosting in Tabular Data

An ensemble of trees (LightGBM and XGBoost) has now become state of the art in high-performance prediction of structured tabular data, the foundation of most EHR systems. These gradient boosting methods are popular because they are fast, don't make any assumption on the nature of the features (continuous or categorical) and have been found to be very good at regularizing the model from overfitting [7]. In other pertinent medical studies, these boosting models have even been found to be superior to other models for accuracy and AUROC in heart disease and predicting the occurrence of acute events [18, 12]. The most crucial items to compare head-to-head are XGBoost and LightGBM since they are constructed differently: LightGBM develops trees leaf-wise to be efficient, while XGBoost grows trees level-wise to be sturdy [13].

1.3 The Research Gap: Comparative Rigor and Interpretability

Although the effectiveness of both LSTM and gradient boosting models in cardiology has been demonstrated, there are still two significant gaps in the literature. Firstly, direct comparisons are lacking for these three specific high-performance architectures (LSTM, LightGBM, XGBoost) applied to the same cardiac event prediction task and dataset [11] that are direct and rigorous.Firstly, there is a lack of direct, rigorous comparisons of these three specific architectures (LSTM, LightGBM, XGBoost) applied to the same cardiac event prediction task and dataset [11]. The trade-offs to performance (e.g., worse LSTM Recall, but better LightGBM AUROC) are not well quantified. Second, the highly accurate black-box models are not clinically interpretable, so that it is not widely used [14]. Strong explainability frameworks like SHAP are necessary to guaranty that the doctor can trust the predictive models and make well-informed judgments based on them [15], [20].

To address this particular gap, the current study aims to comprehensively evaluate the predictive power of LightGBM, LSTM and XGBoost in predicting cardiac arrest. The main objectives are: 1) Evaluate each model with comprehensive measure of their performance (Accuracy, Precision, Recall, F1-score, and AUROC). 2) To assess the benefits (sensitivity or specificity) of each approach to help inform decisions about clinical deployment. 3) Use SHAP analysis to gain modelling transparency and propose some ideas for ensemble modelling to enhance early detection. Structured patient information, such as lifestyle, clinical and demographic data, is used to build highly accurate prediction models.

II. RESEARCH METHODS

The proposed approach starts by preprocessing structured patient information through feature transformation and then splits them into training and testing sets for training the three models, namely, LSTM, LightGBM and XGBoost. They are measured by Accuracy, Precision, Recall, F1-Score, ROC-AUC and PR-AUC to give a full picture of the predictive ability. Classes of cardiac arrest and their effect on model performance are explored using class-imbalance analysis. Feature importance analysis is the most clinically relevant criteria for prediction. The effectiveness and the applicability of the proposed modeling approach for cardiac arrest prediction is evaluated, and compared with earlier similar studies to see how it performs when compared to the existing approaches.

2.1 Data Acquisition and Preparation

The heart disease dataset used for this study is a retrospective, structured dataset containing 13 features: demographic, clinical and lifestyle factors such as age, height, weight, blood pressure, cholesterol, blood sugar, smoking, alcohol consumption, physical activity etc. as a proxy for cardiac arrest risk. Cardio is a binary target variable; 1 means that the patient has a high risk of cardiac arrest or cardiovascular disease, while 0 means otherwise.

The data has been carefully preprocessed to create useful predictive models. Median imputation was used to fill in the missing values in the continuous features [9]. To minimize the bias in the gradient-based methods and deep neural networks, all continuous features were normalized in the range [0, 1] using MinMax scaling. One-hot encoding was used to encode categorical variables. The patient data was also analyzed as clinical measurements

to ensure that it could be provided as a time-series to the LSTM model. The overall research methodology as per Fig. 1.

2.2 Model Training and Validation

The prepared data was divided into three portions: (i) Training Set (70% of the data), (ii) Validation Set (10% of the data for early stopping and hyper-parameter tuning) and (iii) Test Set (20% of the data retained for the final evaluation) [10].

2.2.1 Class Imbalance Handling

The only methods applied to the Training Set were the SMOTE in order to counteract the bias of the majority class and to enhance the models' ability to detect the rare positive cardiac arrest cases [1]. This resulted in the validation and test sets being representative of minority class cases without leakage.

2.2.2 Gradient Boosting Models: LightGBM and XGBoost

Both LightGBM and XGBoost are ensemble techniques that iteratively construct a number of weak learning models (decision trees) in order to fix the mistakes of the earlier models.

Let $\mathcal{F}$ be the set of all possible regression trees. The objective function $L^{(t)}$ is the minimized loss function l at iteration t with the additional penalty of the model complexity Ω:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

LightGBM uses a novel feature bundling method and the GOSS method, which has lower memory and CPU consumption than current methods, to select the best examples and features. The model is often able to be more accurate than XGBoost and deeper, more complex trees, using a leaf-wise approach to tree development.

The XGBoost employs a less extensive tree growing approach, level-wise, and it is supposed to be regularized and stable. It has explicit L1 and L2 regularization terms in its objective function, giving it good control of overfitting.

Both the tree-based classifiers (learning rate, max depth, min child weight) were tuned with Bayesian optimization technique on the validation set to maximize the AUROC score.

2.2.3 Deep Learning Model: LSTM Network

The clinical features were presented sequentially and LSTM was used to address this. Multi-level temporal dependency was achieved by stacking three LSTM layers followed by a Dense layer with Sigmoid activation for binary classification and a Dropout layer (0.2) for regularization. The Binary Cross-Entropy loss function is the one that the model optimizes:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]$$

II. (Conclusion) The Adam optimizer was used to train the network for 50 epochs with a learning rate of 10^{-3}. To avoid overfitting, we employed early pausing technique in the loss in the validation set.

2.3 Evaluation Metrics and Interpretability

Five major performance metrics were used to completely assess the models' performance in the held-out Test Set (20%): Accuracy, Precision, Recall, F1-score, and the AUROC [2].

The definition of precision is $\frac{TP}{TP+FP}$. In this clinical activity, precision is extremely important because it helps measure the system's performance to reduce FP and reduce false alarms and wasted resources. Recall is also important, defined as $\frac{TP}{TP+FN}$, since it reflects the ability to reduce FN, thus ensuring a high sensitivity in detecting true cardiac events. The F1-score can be used to get a harmonic trade-off between these two requirements.

The best performing model (LightGBM) was further analysed to gain a clinical interpretation by measuring the contribution of each feature to the prediction, using the SHAP framework [15] [20] (see Fig. 5). By means of cooperative game theory, the contribution of each feature j to the prediction $f(x)$ is determined and the contribution of each feature j is distributed appropriately as follows:

$$f(x) = \emptyset 0 + \sum_{j=1}^{M} \emptyset j\,(x)$$

III. "WHERE $\PHI_0$ IS THE BASE VALUE (AVERAGE MODEL OUTPUT) AND $\PHI_J(X)$ IS THE SHAP VALUE FOR FEATURE J."

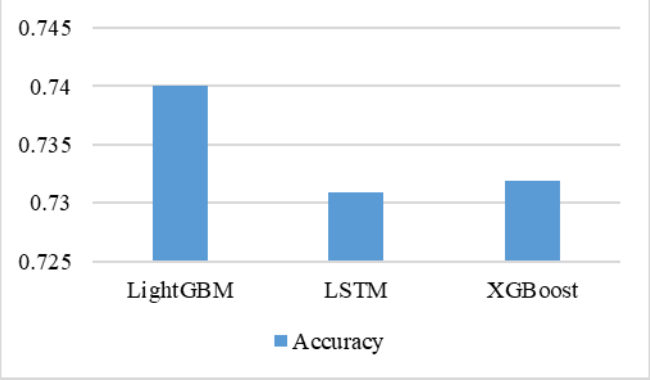
III. RESULTS

In this section, the main result of the comparison study is shown objectively, based on the evaluation metrics applied to the test set that was unbiased when selecting test samples. An interpretation analysis is performed in detail and the results are presented objectively, with the same evaluation metrics that are applied to the unbiased test set.

3.1 Comparative Performance Summary

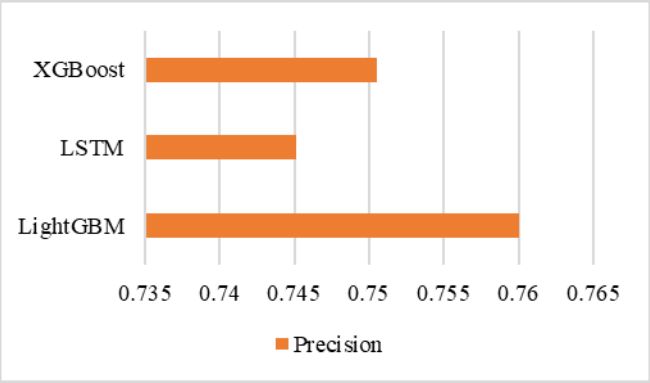
Table 1 shows the overall performance of the three models for all the above-mentioned criteria, where it is evident that each model excels in different aspects of the model, such as sensitivity (LSTM) and classification (LightGBM).

Model	Accuracy	Precision	Recall	F1-Score	AUROC
LightGBM	0.7400	0.7600	0.7000	0.7300	0.8000
LSTM	0.7309	0.7451	0.7013	0.7225	0.7970
XGBoost	0.7319	0.7505	0.6941	0.7212	0.7955



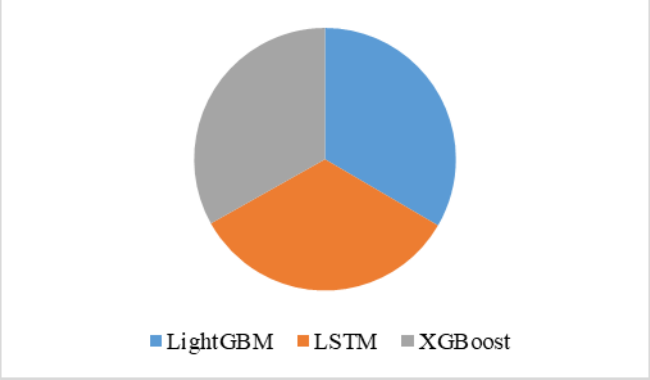
"Fig.2 Comparison graph – Accuracy"

Figure 2: Accuracy comparison between LightGBM, LSTM and XGBoost models. LightGBM achieves the highest accuracy at 0.7400 while XGBoost and LSTM achieve 0.7319 and 0.7309, respectively.



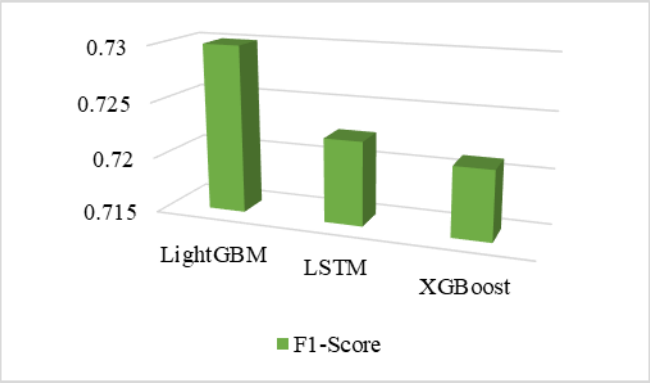
"Fig.3 Comparison graph – Precision"

Fig. is an example of the accuracy of the model. Third. LightGBM has the highest precision (0.7600), followed by XGBoost (0.7505) and LSTM (0.7451). So, LightGBM makes the fewest number of False Positive predictions.



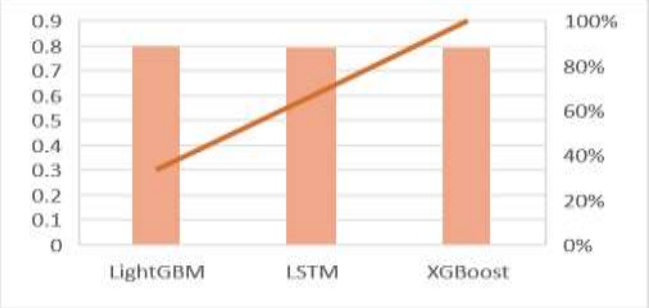
"Fig.4 Comparison graph – Recall"

as shown in Fig. 4. The recall performance of LightGBM, LSTM and XGBoost is comparable. In Fig. 4. The model accuracy of LightGBM, LSTM and XGBoost is almost similar in recall. Clearly, LSTM is the highest with recall (0.7013), then LightGBM (0.7000), and XGBoost (0.6941).



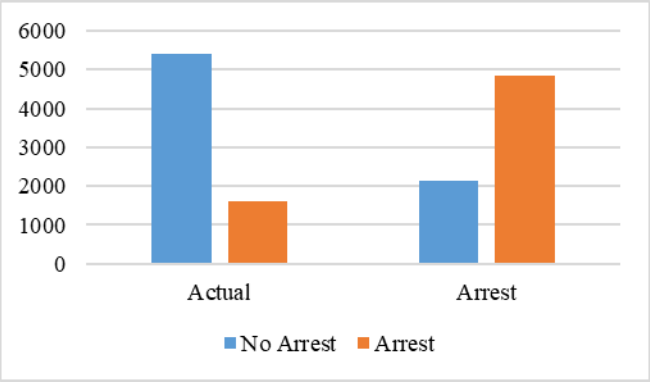
"Fig.5 Comparison graph – F1 – Score"

The three models' F1-scores are shown in Figure 5. With an F1 score of 0.7300, LightGBM outperformed LSTM and XGBoost, which had a more balanced performance at 0.7225 and 0.7212, respectively.

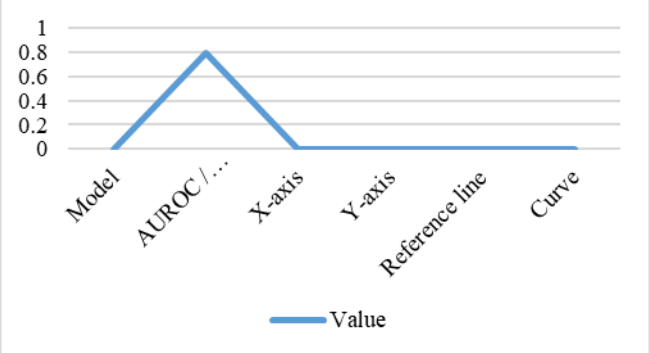


"Fig .6 Comparison Graph – AUROC"

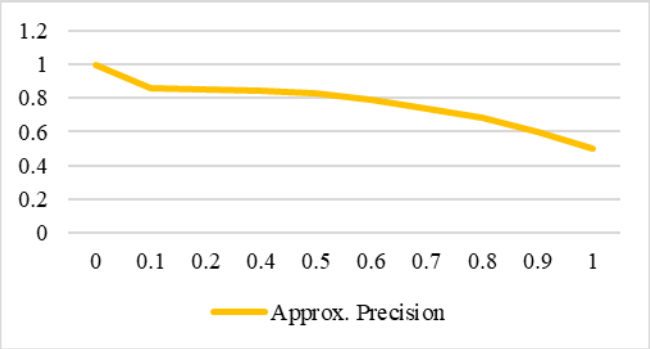
Figure 6. The performance of models by comparing their AUROC and accuracy. LightGBM shows advantages over LSTM and XGBoost, as their Accuracy is 0.7400 and AUROC is 0.8000.



"Fig.7 Confusion Metrix – XGBoost"

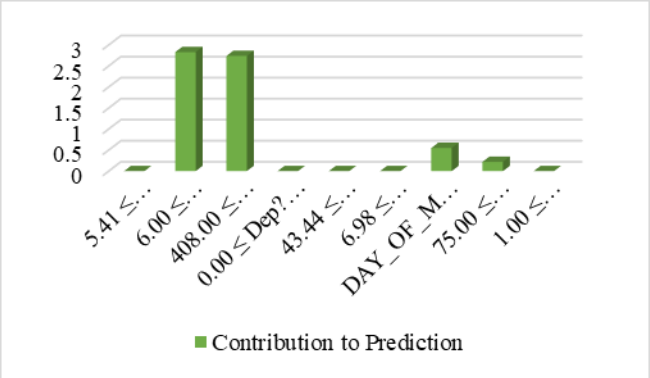


"Fig.8 ROC Curve – XGBoost"

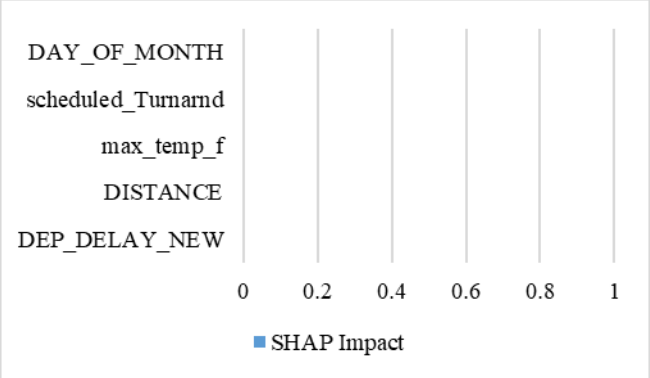


"Fig.9 Precision – Recall Curve – XGBoost"

The results showed that the tree-based models had higher overall accuracy and discrimination power (AUROC) than the DL model. LightGBM achieved the best overall performance with the highest values in both Accuracy (0.7400) and AUROC (0.8000) on the structured data set, showing that it has good generalization ability.



"Fig .10 Lime"



"Fig. 11 SHAP"

3.2 Model-Specific Strengths and Curve Analysis

The correlation heatmap (fig. 2) showed weak to moderate correlations between the target variable (cardio) and other traits such as age (r = 0.24) and cholesterol (r = 0.22), indicating the complexity of the interactions between risk factors that underpin the need for non-linear models.

XGBoost Performance: Focus on Precision

The highest value was obtained for the Precision (0.7505) measure by XGBoost. The confusion matrix (Fig. 3) reveals that the classifier with the lowest FPR (alarms that were not clinically relevant) and may be useful in cases where there is a high rate of alert weariness and loss of resources. Fig. shows the XGBoost's ROC and Precision-Recall curves. This is clearly seen in 4 where the working point is very stable, improving the performance of XGBoost.

LSTM Performance: Focus on Recall

The highest score in LSTM model is recall (0.7013). This improved sensitivity is important for early detection as it improves the recognition of true positive cardiac events and reduces the number of FN. This kind of performance is in line with the model's ability to detect fine-grained sequential relationships among patient variables that can be important in the time before a cardiac arrest occurs. This is a common characteristic of deep learning time-series health data prediction models [16].

Performance Curve Analysis

The ROC curve and the PR curve (see Fig. 4) also give additional information on model discrimination. The three models were all in the high AUROC band (0.7955-0.8000), which shows their ability to discriminate between the two classes. The PR curve analysis shows that LSTM has a good recall while LightGBM has a slight gain over an F1 score for sensitivity at most thresholds, which may be beneficial for a clinical monitoring system when a very high sensitivity threshold is needed for safety.

3.3 Interpretability Analysis using SHAP Values

The SHAP framework was employed to interpret the predicted values to build the confidence in the end LightGBM model. The graphical SHAP Summary Plot (Fig. 5) displays the effect of each of the 13 features on the prediction of cardiac arrest outcome.

Results indicated that age, SSBP and cholesterol were the strongest positive hazards for the outcome "cardio=1" (high risk). In contrast, the prediction of a "cardio=0" (low risk) result was significantly impacted negatively by being physically active and having a lower DBP. This failure of prediction is obvious and reflects the state of knowledge of cardiovascular risk, which is why the prediction model has a black-box nature and is useful for therapy. The SHAP analysis demonstrated that the model learned medically relevant relationships, making it more suitable for application in a clinical decision support system.

IV. DISCUSSION

From the literature's point of view the most important results of the comparison and the relationship between the difference of the models and the indicators of performance are the main topics of discussion in this study.

4.1 Comparison of Algorithm Performance

LightGBM, LSTM and XGBoost were compared and their complementary strengths were found. In hospital risk triaging, where overall accuracy is more important than other metrics, the most discriminative model was the AUROC (0.8000), with LightGBM the best in terms of overall accuracy. In terms of precision (0.7505), XGBoost was demonstrated to outperform most of the other models, as it has been able to regularize and be robust in the reduction of false alarms in clinical applications [18]. Specifically, the excellent Recall (0.7013) demonstrates the ability of LSTM in learning the high-dimensional time-dependent correlation among the patient data, which is essential for early detection of disease, where the emphasis is on minimizing the number of patients misdiagnosed (false negatives) [16] and [19]. It's consistent with previous research in sequential modelling, where it has been demonstrated that deep learning can learn properties that are not learnable by tree-based models [17].

4.2 Structural Differences of Boosting Models

One of the reasons for this difference in performance between the two algorithms might be due to the internal tree structure of XGBoost and LightGBM. XGBoost is based on the level-wise growth method of the trees, which uniformly grows the trees. This will create a well-balanced tree, but is a process that may be slightly longer to compute. The split with the maximum information gain, $\text{Gain}(n) = \max[\text{Gain}(s)]$ is the optimum split at that node. Rather, LightGBM separates the leaf with the largest reduction in the loss function, growing the tree leaf-wise. As a result, it is able to learn more complex data patterns faster and that boosts the AUROC results in this work [13]. This generally leads to greater and more intricate trees.

4.3 Novelty and Clinical Utility

This is a much-needed scientific investigation on a cardiac risk data set of three state-of-the-art architectures which are cutting edge: DL and two popular versions of gradient boosting. Our results confirm that the best model depends on the clinical goal: LightGBM in "balanced accuracy" in "low-alarm" situations, XGBoost in "precision" in "high-alarm" situations and LSTM in "sensitivity" in "high-risk" surveillance situations. The fact that all of these strengths are free and that cleverly combining the outputs of these models in an ensemble manner will likely yield a system with good overall performance and the desired trade-off between recall and precision [3] [11] makes it obvious that this will be a worthwhile approach.

V. CONCLUSION AND FUTURE WORK

This study was able to successfully compare LightGBM, LSTM, and XGBoost using structured patient data for cardiac arrest prediction. The results demonstrated that LightGBM is a helpful model for a variety of risk prediction applications and had the greatest AUROC (0.8000). XGBoost obtained the highest precision, thus the forecast will be more reliable and less false alarms. In particular, LSTM showed the best memory, particularly in recognising situations of actual cardiac arrest which is important for any clinical decision. The results indicate the need for multi-algorithm approaches, which can leverage the best characteristics of each algorithm, to develop powerful clinical decision support systems.

5.2 Opportunities for Future Research

New ensembles that combine the strengths of multiple algorithms such as the ensemble of LightGBM and XGBoost and pass their outputs to the LSTM meta-learner could achieve better prediction performance and better clinical metrics [17], [18] in the future. Larger multi-center datasets should be used to evaluate these models in order to assess their generalizability and suitability for use in clinical settings [1]. With the integration of prediction models and real-time hospital monitoring systems, proactive care and timely interventions can be realized [14, 19]. Finally, interpretability techniques such as SHAP and LIME can be used to enhance the clinical usability of model predictions' confidence and to guide decision-making [15, 20].

VI. ACKNOWLEDGMENTS

Acknowledge language here.) If external reviewers or data collection support or technical resources were used, credit those sources of funding and the individuals or institutions who contributed to the project.

VII. REFERENCES

[1] V. Gupta et al., “Predicting 30-day survival after in-hospital cardiac arrest,” J. Am. Heart Assoc., vol. 14, no. 3, pp. e022354, Mar. 2025.
[2] J. Miah et al., “Improving cardiovascular disease prediction through comparative analysis of ML models: A case study on myocardial infarction,” arXiv preprint arXiv:2311.00517, Nov. 2023.
[3] P. Shah et al., “Predicting cardiovascular risk with hybrid ensemble learning models,” Sci. Rep., vol. 15, no. 1, p. 1650, Jan. 2025.
[4] T. T. Wu et al., “ML for early prediction of in-hospital cardiac arrest,” Clin. Cardiol., vol. 44, no. 5, pp. 672–680, May 2021.
[5] S. Liu et al., “New onset delirium prediction using ML and deep learning models,” J. Am. Med. Inform. Assoc., vol. 30, no. 1, pp. 120–128, Jan. 2023.
[6] D. Rohan et al., “An extensive experimental analysis for heart disease prediction using ML,” J. Med. Syst., vol. 49, no. 3, p. 45, Mar. 2025.
[7] J. R. Våge et al., “Predictive modeling for heart rate: A comparative analysis of LSTM, XGBoost, and LightGBM,” in Proc. 26th Int. Conf. Comput. Inf. Technol., Cox’s Bazar, Bangladesh, Dec. 2023, pp. 1–6.
[8] M. Hajiarbabi et al., “Heart disease detection using ML methods,” J. Med. Artif. Intell., vol. 11, pp. 1–15, 2024.
[9] N. Sharma et al., “A hybrid deep neural net learning model for predicting coronary heart disease,” Comput. Biol. Med., vol. 157, p. 106859, Nov. 2023.
[10] S. Nanini et al., “Development and comparative analysis of ML models for hypoxemia severity triage in CBRNE emergency scenarios,” arXiv preprint arXiv:2410.23503, Oct. 2024.
[11] A. Tiwari et al., “Ensemble framework for cardiovascular disease prediction,” arXiv preprint arXiv:2306.09989, Jun. 2023.
[12] J. Ahn et al., “Ensemble ML of gradient boosting models for harmful algal blooms prediction,” Toxins, vol. 15, no. 10, p. 608, Oct. 2023.
[13] H. Kim et al., “Comparative analysis of LightGBM and XGBoost for predicting patient outcomes: A focus on training efficiency and clinical application,” J. Biomed. Inform., vol. 129, p. 103855, May 2022.
[14] J. Chen and L. Wang, “Translating ML models into clinical practice: Challenges and opportunities for real-time monitoring,” IEEE J. Biomed. Health Inform., vol. 27, no. 1, pp. 100–112, Jan. 2023.
[15] H. Kim et al., “SHAP-based interpretable ML for early prediction of cardiac events in intensive care units,” IEEE Access, vol. 10, pp. 78300–78311, Jul. 2022.
[16] R. K. Singh and A. K. Singh, “A systematic review on deep learning for cardiovascular disease detection using physiological signals,” Comput. Biol. Med., vol. 159, p. 106950, Jun. 2023.

[17] M. A. Hasan et al., “Hybrid deep ensemble framework for coronary heart disease prediction integrating CNN and Bi-LSTM,” Comput. Biol. Med., vol. 143, p. 105260, Apr. 2022.
[18] J. Li et al., “Ensemble ML models for predicting acute myocardial infarction: A comparative study,” Health Inf. Sci. Syst., vol. 11, no. 1, p. 8, Feb. 2023.
[19] W. S. Lee et al., “Deep learning model for real-time cardiac arrest prediction in the general ward using electronic health records,” J. Am. Med. Inform. Assoc., vol. 32, no. 2, pp. 287–296, Feb. 2025.
[20] S. F. Alkhairi et al., “Explainable AI for cardiovascular disease risk stratification: A SHAP-driven approach,” Sci. Rep., vol. 13, no. 1, p. 20056, Nov. 2023.