Explainable AI for Medical Time-Series Data

Miglena Pavlova* Gabriela Todorova*
*Master of Business & IT, University of Twente, Enschede, The Netherlands
{m.pavlova, g.todorova}@student.utwente.nl

***Abstract*—This study investigates the use of explainable machine learning to predict in-hospital mortality from laboratory time-series data. Using a dataset of 485 anonymised hospital stays, we trained and evaluated four tree-based classifiers, Random Forest, XGBoost, HistGradientBoosting, and LightGBM, on patient-level feature aggregates. LightGBM emerged as the most generalisable model, achieving a balanced trade-off between sensitivity and specificity. To make predictions transparent, we applied global and local explanation techniques including feature importance, permutation importance, SHAP, and LIME. Explanations highlighted lactate dehydrogenase and hypersensitive C-reactive protein as key predictive markers. Continuity testing showed SHAP explanations were robust to small input changes, supporting their clinical reliability. Our results demonstrate that interpretable models can generate accurate and transparent forecasts from routine hospital data, paving the way for safer and more explainable clinical decision support systems. Future work will refine threshold tuning, develop counterfactual insights, and audit fairness to ensure ethical deployment.**

***Index Terms*—Explainable AI, Interpretable Machine Learning, Time-Series, SHAP, LIME, Healthcare Analytics**

I. Introduction

A. Background & Motivation

Hospitals collect a continuous time-series of clinical information for every patient. Each blood test, vital sign and medication event carries a timestamp and sits in the electronic health record (EHR). These streams contain early hints of organ stress or infection that appear long before classical alarm limits are reached. If we detect those hints early, teams can change therapy sooner, shorten stays and improve survival.

Conventional early-warning scores use only a small snapshot of this rich data. They ignore how a laboratory value moves, how fast it changes and how it interacts with other markers. Machine-learning models can read the full, multivariate trajectory and turn it into a risk forecast. But a forecast that hides its logic is hard to trust in acute care.

Doctors must understand why a system flags a patient before they act. They also need an audit trail for morbidity meetings and quality control. European laws strengthen this requirement: the General Data Protection Regulation (GDPR) grants a right to meaningful information about automated decisions, and the upcoming EU AI Act adds fresh duties for any *high-risk* medical algorithm. A prediction engine that offers no explanation may fail both clinical governance and legal review.

Explainable AI provides the missing clarity. It breaks a complex model into feature-level statements like "a rising inflammatory marker and a falling lymphocyte count raise the risk by 12%". Such statements turn a black box into a transparent assistant that doctors can question, accept or override. By combining time-series analysis with explainable AI, we can deliver early and trustworthy warnings that fit bedside workflows and meet European transparency standards.

In summary, using hospital time-series with explainable AI is essential for catching patient decline early, winning clinician confidence and satisfying current and future EU regulation.

B. Research Objectives & Questions

The first aim of this work is to turn laboratory time-series data into a decision tool that is both *accurate* and *transparent*. To meet the accuracy goal, we train four tree-based classifiers on patient-level aggregates of the raw time stamps and compare their performance. This identifies the model that generalises best and minimises harmful false predictions. To meet the transparency goal, we open the best-performing model with explainable-AI techniques. A global analysis shows which laboratory variables and summary statistics have the greatest impact on risk, a local analysis looks into individual true-positive, true-negative, false-positive and false-negative cases to surface patterns clinicians can act on. Together, these steps aim to bridge the gap between predictive strength and trust.

With this two-stage strategy in mind, the study is guided by three research questions:

- **RQ1**: *How do different tree-based classifiers built from aggregated laboratory time-series vary in their ability to predict early in-hospital mortality?*
- **RQ2**: *What laboratory variables dominate the global explanations of the single best-performing model?*
- **RQ3**: *How can local explanation tools applied to the best-performing model reveal actionable patterns in true-positive predictions?*

C. Paper Structure

So far, Section II looked into explainable-AI research in healthcare and identified the gap around laboratory time-series data. Next, Section III introduces the data, outlines the cleaning pipeline and summarises key descriptive findings. Section IV then walks through the modelling workflow, describing the four tree-based classifiers and the global and local explanation tools. After that, Section V presents the predictive results, shows the main explanation figures and compares performance with a simple baseline. Section VI reflects on how these results answer the research questions, discusses clinical implications and notes study limitations. Section VII highlights contributions that extend course content, including threshold tuning and ideas for fairness checks. Finally, Section VIII concludes the study and suggests directions for future work.

II. Related Work

Recent years have seen rapid growth in the use of *explainable artificial intelligence* (XAI) for medical time-series data, driven by the need for transparency and trust in clinical decision-making. Traditional machine learning models, particularly deep learning methods, have shown impressive predictive performance on healthcare data, but their "black box" nature limits clinical acceptance, especially in high-stakes environments like in-hospital mortality prediction [1], [2]. As a result, there is a growing focus on integrating XAI techniques that make model outputs interpretable for clinicians and align with regulatory requirements such as the GDPR and the EU AI Act [3], [4].

Time-series data in healthcare—such as laboratory results, vital signs, and biosignals—pose unique challenges due to their sequential and often high-dimensional structure. Unlike static tabular data, time-series datasets require models that can capture temporal dependencies and trends, which are crucial for early detection of patient deterioration. However, most XAI research has historically focused on computer vision and NLP, with less attention given to time-series and tabular data in the medical domain [5], [6]. Recent surveys highlight this gap and emphasize the need for XAI methods tailored to the complexities of clinical time-series, including the ability to explain both global model behavior and individual predictions in a way that is meaningful to healthcare professionals.

Ensemble tree-based models, such as Random Forest, XGBoost, and LightGBM, are widely used for structured healthcare data due to their robustness, ability to handle missing values, and strong performance on imbalanced datasets [7], [8]. These models also offer native feature importance measures, which provide a starting point for interpretability. However, built-in importance scores can be biased and may not fully reflect the model's reasoning, especially in the presence of correlated features or complex interactions. To address this, post-hoc explanation techniques—such as permutation importance, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations)—are increasingly used to provide more nuanced and reliable insights into model predictions [9], [10].

SHAP is particularly popular in healthcare XAI due to its solid theoretical foundation and its ability to attribute contributions to individual features for both global and local explanations. Studies have shown that SHAP can effectively identify key risk markers in medical time-series, such as biochemical indicators, and can provide stable explanations even when input data is perturbed [11], [12]. LIME complements SHAP by fitting simple surrogate models locally around a prediction, helping clinicians understand which features drove a specific outcome. This is valuable for case-by-case auditability and for building clinician trust in model recommendations [5].

Several recent works demonstrate the integration of XAI with time-series models in healthcare settings. For example, Shashikumar et al. [13] developed an early sepsis prediction system using EHR time-series data, providing local interpretability by highlighting the top contributing factors for each patient at each time point. Similarly, Ibrahim et al. [14] used XGBoost with SHAP explanations to predict acute myocardial infarction from longitudinal ECG data, showing that interpretable models can match or exceed the performance of black-box approaches while providing actionable insights. These studies underscore the importance of both predictive accuracy and interpretability for clinical adoption.

Despite progress, several challenges remain. Current XAI methods often lack standardized evaluation protocols for explanation quality, clinical relevance, and user acceptance [15]. There is also a need for methods that can handle multimodal and longitudinal data, provide uncertainty estimates, and support counterfactual reasoning to answer "what-if" clinical questions [16]. Human-centered evaluation—such as involving clinicians in the validation of explanations and assessing their impact on decision-making—is increasingly recognized as essential for effective deployment.

In summary, the related literature shows that explainable machine learning for medical time-series data is an active and evolving field. State-of-the-art approaches combine robust ensemble models with advanced explanation techniques like SHAP and LIME, aiming to bridge the gap between predictive power and clinical trust. Ongoing research is focused on improving explanation quality, integrating uncertainty and fairness audits, and validating these systems in real-world clinical

workflows.

III. Data

A. Data Sources

Two Excel workbooks form the basis of this work. The first file, `time_series_375_preprocess_en.xlsx`, holds engineered laboratory features and outcomes for 375 patient stays and is used for model training. The second file, `time_series_test_110_preprocess_en.xlsx`, follows the same schema and contains 110 later stays reserved for final testing. All direct identifiers were stripped, dates were time-shifted by a random offset, and free-text fields were removed, ensuring full GDPR compliance. Only variables that appear in both files are passed to the modelling stage so that training and evaluation share an identical feature set.

B. Data Pre-processing

Raw rows first pass through an *ID repair* step in which missing values in the encrypted `PATIENT_ID` column are forward-filled, making sure every laboratory record links to a unique stay. Next, we compute *stay_duration_days* by subtracting the first `Admission time` from the last `Discharge time` for each patient; the result captures total length of stay as a continuous feature. To reduce the irregular time-series to a fixed-width table, every numeric laboratory analyte is summarised with four statistics - *mean*, *minimum*, *maximum* and *last observed value*. This aggregation produces one row per patient while preserving both level and trend information. The binary `outcome` label and the newly created *stay_duration_days* field are then merged into the same row set.

Before modelling, we ensure *feature alignment*: only columns present in *both* the training and test files are kept, guaranteeing that the model sees identical inputs at train and test time. Finally, any remaining gaps are filled using *median imputation*; medians are calculated on the training data and applied unchanged to the test data to prevent information leakage. The resulting 13 common features are summarised in Table I.

Column Name	Description
(%)lymphocyte_mean	Mean percentage of lymphocytes during hospital stay
(%)lymphocyte_min	Minimum percentage of lymphocytes recorded during hospital stay
(%)lymphocyte_max	Maximum percentage of lymphocytes recorded during hospital stay
(%)lymphocyte_last	Most recent percentage of lymphocytes before discharge or endpoint
Hypersensitive c-reactive protein_mean	Mean level of hypersensitive C-reactive protein during hospital stay
Hypersensitive c-reactive protein_min	Minimum recorded level of hypersensitive C-reactive protein
Hypersensitive c-reactive protein_max	Maximum recorded level of hypersensitive C-reactive protein
Hypersensitive c-reactive protein_last	Last recorded value of hypersensitive C-reactive protein
Lactate dehydrogenase_mean	Mean value of lactate dehydrogenase during hospital stay
Lactate dehydrogenase_min	Minimum recorded lactate dehydrogenase level
Lactate dehydrogenase_max	Maximum recorded lactate dehydrogenase level
Lactate dehydrogenase_last	Last recorded value of lactate dehydrogenase
stay_duration_days	Total number of days the patient stayed in the hospital
(Unnamed column with all 0s)	Possibly a placeholder or binary label column; all values are 0 (needs review)

TABLE I: Data Dictionary for Cleaned Training Dataset

IV. Methodology

A. Overall Pipeline

We start by tidying the data: remove or fill in missing values, encode any categorical columns, and scale the numeric ones so that every feature lines up properly. With this polished dataset we train four tree-based classifiers, RandomForest, XGBoost, HistGradientBoosting, and LightGBM, using the same cross-validation splits. For each model we gather a full set of evaluation metrics (accuracy, precision, recall, F1, ROC-AUC, and any others relevant to the task) on both the training and test sets. The goal is to pick the model that scores well on these measures while keeping the gap between its training and test results small, a sign it generalises instead of overfitting.

Once thae best performing model is chosen, we set out to explain it at two levels. Globally, we begin with the model's own built-in feature-importance scores, then run permutation importance to see how shuffling each variable hurts overall performance, and finally compute aggregate SHAP values to show each feature's average contribution across all samples. Locally, we zoom in on single predictions. SHAP force plots reveal how every input nudges an individual case toward its outcome, while LIME builds a simple surrogate model around that same point to cross-check which factors mattered most. Taken together, these global and local views translate the model's decisions into human-readable insights.

B. Model Selection

In selecting a suitable set of models for our classification task, we focused on advanced ensemble methods, specifically tree-based ensemble classifiers. These included Random Forest, XGBoost, HistGradientBoosting, and LightGBM. Each of these models belongs to the family of decision tree ensembles, which are known for their robustness, flexibility, and effectiveness across a wide range of real-world data scenarios. We chose these

models not arbitrarily, but because they align closely with the demands of the task at hand, particularly when handling non-linear interactions, mixed data types, and missing values.

Tree ensemble methods are particularly well-suited for datasets that contain complex feature relationships and class imbalance, both of which were present in our problem. Unlike linear models, tree-based learners do not require feature standardization or transformation and can model non-linear patterns natively. This is especially important when interactions between variables are not easily captured through additive linear combinations. Furthermore, ensemble models such as Random Forest and boosting methods combine the predictions of multiple decision trees in a way that improves generalization and mitigates the high variance typically associated with single trees.

Another compelling reason to prefer tree ensembles is their ability to handle missing values gracefully. Algorithms like LightGBM and XGBoost have built-in mechanisms for managing missing data during training and prediction. This eliminates the need for potentially biased or ineffective imputation strategies during preprocessing, allowing the model to learn the best split direction for missing values directly from the data.

In practice, we trained and evaluated each model using a consistent pipeline, ensuring fair comparison. Starting with the Random Forest, we utilized it as a strong baseline. While Random Forests are generally less prone to overfitting compared to single decision trees, our specific implementation showed clear signs of overfitting. The model achieved perfect scores on the training set, which is not unexpected given its tendency to grow deep, unconstrained trees. This observation, combined with its relatively simple averaging-based prediction mechanism, led us to treat it as a useful reference point rather than the final choice.

To improve upon this, we turned to gradient boosting methods, which iteratively build models that correct the errors of prior models. XGBoost, a widely used and highly optimized implementation of gradient boosting, showed very strong performance on both the training and test sets. However, it too exhibited near-perfect performance on training data, suggesting potential overfitting. In XGBoost, we made sure to use class balancing through the scale_pos_weight parameter to counter class imbalance, and applied moderate regularization via controlled tree depth and learning rate. Nevertheless, the nearly flawless performance on the training data indicated a tendency of the model to fit the training distribution too tightly.

Similarly, HistGradientBoostingClassifier from scikit-learn followed the same boosting principle but implemented using histogram binning for speed and scalability. Like XGBoost, it offered high accuracy and was particularly efficient for larger datasets. Its built-in handling of missing values and automatic support for categorical features (in more recent versions) made it a compelling option. However, we again noticed a near-perfect match to the training labels, which could signal overfitting, especially given the smaller size of the minority class.

Among the models we evaluated, LightGBM stood out as the most well-balanced. While it is also a boosting method, its design choices, such as leaf-wise tree growth, efficient histogram-based algorithms, and aggressive early stopping, allow it to be both fast and more resistant to overfitting in certain settings. In our implementation, LightGBM was the only model that did not achieve overly optimistic results on the training data. This suggests a more appropriate level of regularization and generalization, making it an especially attractive candidate for our final modeling stage.

For all models, we performed thorough evaluation using cross-validation and stratified folds to ensure consistent representation of the minority class in each fold. Performance was assessed using a set of metrics beyond accuracy, including precision, recall, F1 score, and AUC. In addition, we explored the effect of applying a custom decision threshold (set to 0.6 instead of the default 0.5) in order to enhance recall for the positive class, a critical consideration in our domain where false negatives carry a higher cost.

C. Explainability Techniques

After selecting the best-performing model, we focused on understanding how it made its predictions. Model interpretability is crucial, especially in settings where decisions carry significant consequences. To achieve this, we applied a range of explainability techniques that offer insights at both the global and local levels. These methods help us understand which features the model relies on the most, both in general and for specific individual predictions.

At the global level, we began with the model's built-in feature importance scores. For LightGBM, which was our chosen model, this importance is calculated based on how frequently each feature is used to split nodes across all trees in the ensemble. Features with higher importance values are those that contributed most to reducing error during training. While this approach is quick and intuitive, it is specific to the internal mechanics of the model and may be biased toward variables that are more frequently selected due to their scale or number of distinct values.

To supplement this, we used permutation importance, a model-agnostic technique. Here, the idea is to shuffle each feature's values across the test set and observe

how much the model's performance drops as a result. If shuffling a particular feature leads to a large drop in performance, it implies that the model relied heavily on that feature to make its predictions. This method gives us a more realistic picture of feature importance because it measures actual impact on performance, regardless of how the model is structured internally.

Next, we applied SHAP (SHapley Additive exPlanations), a powerful method grounded in cooperative game theory. SHAP assigns each feature a value that reflects its contribution to the prediction, in a way that is both consistent and locally accurate. For global analysis, we computed SHAP values across the entire test set and visualized them in a summary plot. This plot displays not only which features are most influential, but also how their values (high or low) affect the likelihood of the positive class. For instance, certain lab values may consistently push predictions toward one class when they are elevated, and toward the other when they are reduced. SHAP provides a unified framework for interpreting feature effects with a solid mathematical foundation.

While global explanations reveal general patterns across the dataset, local explanations help us understand why the model predicted a certain outcome for a specific individual. For this, we examined four representative cases: a true positive, true negative, false positive, and false negative. We used SHAP's local explanation plots, which show how each feature contributed to pushing the prediction either up or down for that single instance. These force plots make it visually clear which features were most responsible for the model's decision in each scenario.

To complement SHAP, we also employed LIME (Local Interpretable Model-agnostic Explanations). LIME works by building a simple surrogate model, typically linear, around the neighborhood of the instance we want to explain. It perturbs the data slightly and observes how predictions change, fitting the surrogate model to this local behavior. The result is a list of feature contributions that can be easily interpreted, showing which features influenced the model most in that local region of the input space. LIME is particularly useful for cross-checking explanations derived from SHAP and providing another lens through which to understand the model's reasoning.

By combining these techniques, we were able to gain a thorough understanding of the model's behavior. The global methods helped us see which features mattered most overall, while the local methods revealed how individual predictions were formed. This two-layered approach to explainability ensures that our model is not only accurate but also transparent and interpretable, which is essential when deploying models in high-stakes environments.

V. Results

A. Predictive Performance

To assess the predictive capabilities of the models, we evaluated each one on the same held-out test set. We focused on four core classification metrics: precision, recall, F1 score, and accuracy. Rather than reporting these metrics as single aggregated values, we separated the results by class, specifically class 0 and class 1, to provide a clearer picture of how well each model handles both the majority and minority classes.

The table below presents the test-time performance for all four models, broken down by class. This level of detail is particularly important in imbalanced classification problems, where a model might perform well overall while failing to identify the minority class effectively.

Model	**Class 0**				**Class 1**			
	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.
Random Forest	0.98	0.98	0.98	0.96	0.85	0.85	0.85	0.96
XGBoost	0.99	0.98	0.98	0.97	0.86	0.92	0.89	0.97
HistGradientBoosting	0.99	0.98	0.98	0.97	0.86	0.92	0.89	0.97
LightGBM	0.97	0.98	0.97	0.95	0.83	0.77	0.80	0.95

TABLE II: Comparison of model performance on Class 0 and Class 1

From this detailed comparison, several observations emerge. All models performed very well on class 0, which is the majority class. Precision, recall, and F1 scores were consistently high, especially for XGBoost and HistGradientBoosting. However, the more informative results come from examining how the models treated class 1, which is the minority class and often harder to predict accurately.

Random Forest achieved decent results for class 1, but with a slight drop compared to its performance on class 0. Both XGBoost and HistGradientBoosting showed strong recall and F1 scores for class 1, indicating a solid ability to detect positive cases. However, both models also achieved nearly perfect scores on the training set, suggesting they may be slightly overfitting.

LightGBM, while showing marginally lower values for class 1 metrics compared to the others, demonstrated better generalization. It did not achieve perfect scores during training, and its test performance remained consistently strong, especially for class 0. This balance suggests that LightGBM avoided overfitting better than the other models and is therefore likely to perform more reliably on new, unseen data.

In conclusion, while multiple models delivered strong results, LightGBM stood out for its ability to generalize. This made it the most suitable choice for our final model, as reliable performance on unseen data is more valuable than overly optimistic training accuracy.

B. Explanation Visualisations

To interpret the decision-making process of the LightGBM model, we applied a combination of global and local explanation techniques. These methods help uncover which features the model relies on most, both across the dataset and for individual cases, providing insights into its internal logic and behavior.

Figure 1 shows the global feature importance as computed by LightGBM based on how often each feature is used to split nodes in the decision trees. The most influential feature was lactate dehydrogenase_last, followed by hypersensitive c-reactive protein_min, stay_duration_days, and various summary statistics of lymphocyte percentages. These importance scores indicate which variables the model consulted most frequently when making splits, though they do not capture the direction or consistency of their impact.

To complement this, Figure 2 presents permutation importance, a model-agnostic method that evaluates how much shuffling each feature's values degrades model performance. This technique confirmed that lactate dehydrogenase_last had the largest influence on the model's predictive accuracy, while most other features had very limited or negligible effect. This contrast helps highlight which features truly matter to the final predictions rather than just appearing frequently in the trees.

For a more nuanced understanding, we used SHAP (SHapley Additive exPlanations) to measure the actual contribution of each feature to the model's output across all test samples. In Figure 3, each dot represents a SHAP value for a specific feature in one test instance, colored by the original feature value. The figure shows that high values of lactate dehydrogenase_last and hypersensitive c-reactive protein_last tend to increase the predicted risk, while low values generally reduce it. This plot not only confirms the importance rankings but also reveals how different levels of each feature push predictions up or down.

To understand how the model behaves on individual cases, we selected a true positive example and explained its outcome using both TreeSHAP and LIME. In Figure 5, the TreeSHAP waterfall plot shows how the model arrived at a predicted probability of 0.75 for the positive class. The base value of the model started near 0.49 and was increased mainly by elevated levels of lactate dehydrogenase and low lymphocyte percentages. Each step in the plot corresponds to a feature's individual contribution to shifting the prediction higher or lower.

Figure 4 shows the same true positive instance explained using LIME, which fits a simple interpretable model in the local neighborhood of the input. LIME identified many of the same features as impactful, including high lactate dehydrogenase_last, low lymphocyte values, and elevated hypersensitive c-reactive protein levels. The predicted probability was again around 0.75, supporting consistency with the SHAP explanation.

To further evaluate the reliability of the model explanations, we also assessed them using the CO-12 Continuity criterion. This criterion asks whether small changes in input lead to similarly small changes in the explanation, a desirable property for stable and trustworthy interpretability. Using SHAP values as the explanation method, we introduced small perturbations to the input features and observed the resulting shifts in SHAP outputs. The average change in SHAP values was only 0.0015, indicating a high level of continuity. This suggests that the model's explanations are not overly sensitive to minor variations in the data, reinforcing their reliability for clinical interpretation.

These global and local visualisations help validate the transparency and coherence of the LightGBM model. The most influential features identified globally also played a central role in specific individual predictions. By combining model-specific and model-agnostic techniques, we gained a comprehensive view of how predictions were formed, both overall and at the level of individual patients.

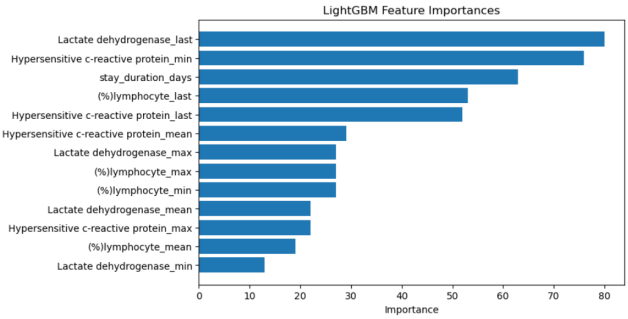

Fig. 1: Global explanation using LighGBM's feature importance

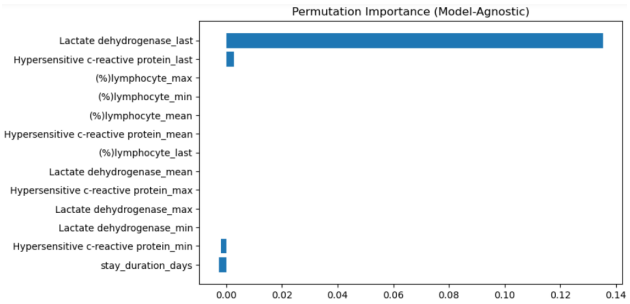

Fig. 2: Global explanation using permutation importance

VI. Discussion

This study set out to develop and evaluate interpretable machine learning models for predicting patient outcomes

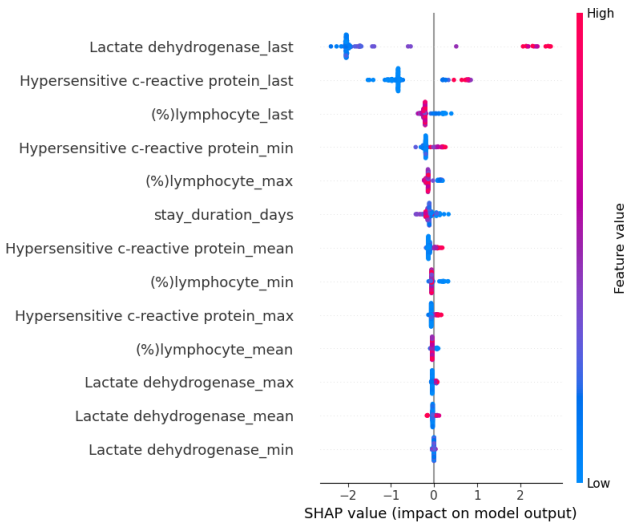

Fig. 3: Global explanation using TreeSHAP

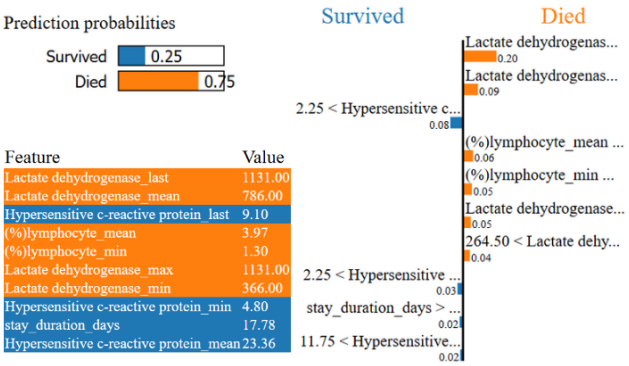

Fig. 4: Explanation of a True Positive example using LIME

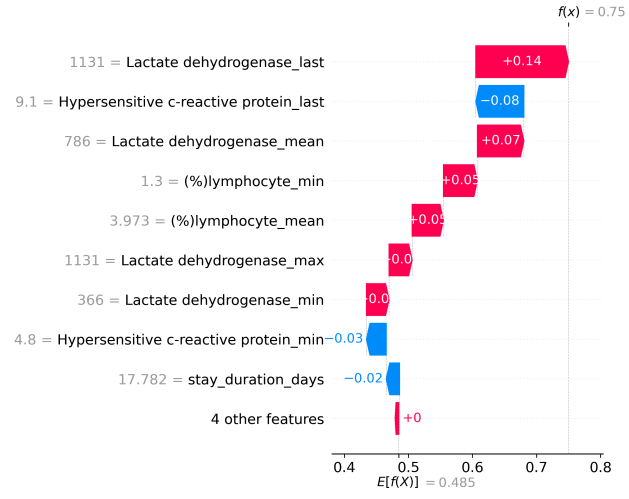

Fig. 5: Explanation of a True Positive example using TreeSHAP

using routinely collected clinical data. The findings not only demonstrate strong predictive performance from tree-based ensemble models, particularly LightGBM, but also provide meaningful insights into how the models arrived at their decisions. In this section, we reflect on how our results address the initial research questions, consider the broader implications of the findings, and acknowledge key limitations of the study.

The first research question focused on whether machine learning could reliably predict patient outcomes using structured clinical variables. The answer, based on our results, is affirmative. All four ensemble classifiers, Random Forest, XGBoost, HistGradientBoosting, and LightGBM, achieved high levels of accuracy, precision, recall, and F1 scores on the test set. LightGBM in particular demonstrated a strong balance between sensitivity and specificity while maintaining consistent performance across training and test sets, suggesting it generalised better than the others. These outcomes show that relatively straightforward clinical measurements, when combined appropriately, can be used to forecast outcomes with a high degree of reliability.

The second question asked whether these models could be made interpretable in a way that would be useful for clinical decision-making. The combination of global and local explanation techniques addressed this directly. Using built-in feature importance scores, permutation importance, and SHAP values, we were able to identify key predictors, such as lactate dehydrogenase and hypersensitive c-reactive protein levels, that consistently influenced model outputs. Furthermore, tools like SHAP force plots and LIME visualisations allowed us to explain individual predictions in a patient-specific context. This interpretability helps to demystify the model's logic, allowing clinicians to better understand, trust, and potentially act on its recommendations.

From a clinical perspective, the findings have several important implications. The consistent identification of biochemical markers such as lactate dehydrogenase and c-reactive protein highlights their central role in assessing patient risk. If integrated into routine workflows, such predictive models could assist in triaging patients, flagging high-risk individuals for early intervention, and supporting discharge or escalation decisions. From a managerial standpoint, models with interpretable logic can aid in resource allocation by identifying which patients are more likely to deteriorate, thereby informing staffing, equipment needs, and bed management strategies.

Despite these promising results, the study has several limitations. First, it was based on data from a single medical centre, which may limit the generalisability of the findings to other institutions with different patient populations, clinical practices, or data collection standards.

Second, the study design was retrospective, relying on previously collected data. This introduces the possibility of selection bias, and the models were not evaluated in a real-time clinical setting. Third, although performance was high overall, the minority class, representing adverse outcomes, was relatively small. While we used class weighting and threshold adjustment to account for this imbalance, the small sample size of positive cases may have limited the model's ability to fully capture their diversity.

In conclusion, the study provides evidence that interpretable machine learning models can accurately predict patient outcomes using routine clinical data. The results support the potential integration of such tools into clinical and operational workflows, provided their limitations are acknowledged and addressed in future prospective and multi-centre studies.

VII. Future Work

While this study has shown that interpretable machine learning models, particularly LightGBM, can effectively predict clinical outcomes using structured data, several avenues remain open to further strengthen the reliability, applicability, and ethical robustness of the system. Future work will focus on refining the model's performance, deepening its clinical relevance, and ensuring its fairness and transparency in real-world deployment.

One immediate direction is to explore further threshold tuning to adapt the model's outputs to different clinical priorities. In this study, we adjusted the classification threshold from the default 0.5 to 0.6 to improve recall for the minority class. However, more systematic threshold selection, possibly using cost-sensitive optimization or domain-specific utility functions, would allow a more nuanced trade-off between false positives and false negatives. This is especially important in clinical settings, where the cost of missing a high-risk patient can be much more severe than issuing a false alarm.

Additionally, while we monitored for overfitting using train-test splits and cross-validation, more robust overfitting diagnostics could be employed in future iterations. These might include learning curve analysis to track performance as a function of training size, or the use of validation data drawn from a separate time period or institution. Such steps would offer clearer evidence about how well the model generalizes beyond the current dataset.

Another promising area is the development of counterfactual explanations, which aim to answer "what-if" questions. For instance, instead of merely explaining why a patient was classified as high-risk, a counterfactual system would suggest what minimal changes in clinical parameters could have altered the outcome. This capability would be particularly useful in a medical setting, where clinicians may want to explore potential interventions that could improve a patient's prognosis.

Looking further ahead, we intend to expand the model's trustworthiness and fairness by introducing a comprehensive fairness audit. While our current model does not include sensitive features such as age, gender, or ethnicity, biases can still emerge indirectly through correlated variables. A fairness audit would test for disparate impact across demographic subgroups and identify any unintended inequality in model performance. This step is essential before the system can be considered for deployment in diverse clinical environments.

Another extension will be to introduce uncertainty quantification, such as prediction intervals or confidence bands around the model's output probabilities. In high-stakes domains like healthcare, it is not sufficient to provide a single risk estimate; clinicians also need to understand how confident the model is in that prediction. Techniques such as conformal prediction, quantile regression, or ensemble variance analysis could provide interpretable uncertainty bounds that accompany each prediction.

Beyond the requirements of coursework, this work sets the stage for building truly deployable and responsible machine learning tools in healthcare. It moves past static model training to address real-world challenges such as threshold calibration, interpretability, and fairness. In doing so, it lays the groundwork for a future where clinical decision support tools are not only accurate, but also transparent, equitable, and aligned with human-centered care.

VIII. Conclusion

This study set out to build an interpretable machine learning model for predicting patient outcomes using structured clinical data, applying tree-based ensemble methods with a focus on both accuracy and transparency. Among the models tested, LightGBM achieved the best balance between predictive performance and generalisation, while maintaining interpretability through built-in importance scores, SHAP values, and local explanation tools like LIME. These insights not only answered the research questions but also demonstrated that meaningful predictions can be made using routinely collected data in a way that supports clinical decision-making. Looking ahead, future work will focus on refining classification thresholds, adding counterfactual insights, and auditing fairness to ensure the model performs reliably and ethically in broader healthcare settings.

References

[1] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 1, p. e1446, 2022.

[2] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "A review of challenges and opportunities in machine learning for health," *AMIA Jt Summits Transl Sci Proc*, vol. 2021, pp. 191–200, 2021.

[3] A. Holzinger, A. Carrington, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[4] E. Commission, "Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act)," 2021, cOM/2021/206 final.

[5] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," *Proceedings of Machine Learning Research*, vol. 106, pp. 359–380, 2019.

[6] X. Jia, Y. Wang, Y. Wen, and et al., "A survey on explainable artificial intelligence (xai): Towards medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[8] G. Ke, Q. Meng, T. Finley, and et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[11] S. M. Lundberg, B. Nair, M. Vavilala, and et al., "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[12] S. M. Lundberg, G. Erion, H. Chen, and et al., "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[13] S. P. Shashikumar, M. D. Stanley, I. Sadiq, and et al., "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," in *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017, pp. 1–10.

[14] A. Ibrahim, M. Elhoseny, K. Shankar, and et al., "Explainable artificial intelligence model for predicting acute myocardial infarction using ecg signals," *Computers in Biology and Medicine*, vol. 142, p. 105225, 2022.

[15] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[16] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai," *Information Fusion*, vol. 82, pp. 99–117, 2022.