Interpreting Machine Learning Predictions with SHAP and LIME for Transparent Decision Making

Stow, May* and Stewart, Ashley Ajumoke**
Department of Computer Science and Informatics, Federal University Otuoke, Nigeria*
Department of Fine Arts and Design, University of Port Harcourt, Nigeria**
maystow@gmail.com*, ashley.stewart@uniport.edu.org**
Orcid ID: https://orcid.org/0009-0006-8653-8363*, https://orcid.org/0009-0006-8425-4236**
DOI: 10.56201/ijcsmt.vol.11.no8.2025.pg22.49

Abstract
Machine learning models increasingly influence critical decisions across diverse domains, yet their complex architectures often operate as black boxes, obscuring the rationale behind predictions and limiting stakeholder trust. This research demonstrates a comprehensive, reproducible workflow for applying explainable artificial intelligence techniques to interpret Random Forest classifier decisions using publicly available data and standard computational resources. The study implements and compares two leading explanation methods, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), on the Titanic survival prediction task to evaluate their consistency and practical utility. The methodology encompasses automated data preprocessing, model training with regularization to prevent overfitting, and systematic generation of both global and local explanations through multiple visualization formats. Results reveal exceptional agreement between explanation methods, with Spearman rank correlation of 0.918 and Pearson correlation of 0.982 for feature importance values. Both techniques consistently identified passenger sex as the dominant predictive feature, contributing approximately 15.4% and 12.5% of model decisions respectively, followed by passenger class and fare. The Random Forest model achieved 84.5% test set ROC-AUC with controlled overfitting (0.040 ROC-AUC gap between training and test sets) while maintaining interpretable complexity through architectural constraints. The implementation executes efficiently on CPU hardware within minutes, eliminating computational barriers to XAI adoption. This work establishes that current explainability techniques can provide reliable, consistent insights into ensemble model decisions while remaining accessible to researchers and practitioners with limited computational resources.

Keywords: *Explainable Artificial Intelligence, SHAP, LIME, Random Forest, Model Interpretability, Machine Learning Transparency*

1. Introduction

The proliferation of machine learning systems in critical decision-making domains has created an urgent need for interpretable and explainable artificial intelligence. As algorithms increasingly determine loan approvals, medical diagnoses, criminal sentencing recommendations, and employment decisions, the ability to understand and validate model reasoning becomes essential for ensuring fairness, accountability, and regulatory compliance (Adadi & Berrada, 2018). The European Union's General Data Protection Regulation (GDPR) explicitly establishes the right to explanation for automated decision-making systems, while similar frameworks emerge globally, mandating that organizations provide meaningful information about the logic involved in algorithmic decisions (Wachter et al., 2017).

Despite significant advances in model performance, the trade-off between predictive accuracy and interpretability remains a fundamental challenge in machine learning deployment. Complex models such as deep neural networks and ensemble methods often achieve superior performance but function as "black boxes," offering little insight into their decision processes (Rudin, 2019a). This opacity creates barriers to adoption in regulated industries, limits debugging capabilities, and prevents stakeholders from identifying potential biases or errors in model reasoning (Lipton, 2018). The explainable AI field has emerged to address these challenges, developing methods that illuminate model behavior while preserving predictive performance.

Recent years have witnessed the development of numerous explainability techniques, broadly categorized into model-agnostic and model-specific approaches. Among model-agnostic methods, SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) have gained widespread adoption due to their theoretical foundations and practical applicability across diverse model types. SHAP leverages cooperative game theory to assign each feature an importance value based on Shapley values, ensuring consistency and local accuracy (Lundberg & Lee, 2017). LIME approximates model behavior locally through interpretable surrogates, optimizing for fidelity in the neighborhood of specific instances (Ribeiro et al., 2016). While both methods have demonstrated effectiveness in various applications, systematic comparisons of their relative strengths, computational requirements, and practical trade-offs remain limited.

Existing research typically evaluates explainability methods in isolation or focuses on specific application domains without providing reproducible workflows that practitioners can adapt. Studies often employ complex models and large-scale datasets that require substantial computational resources, creating barriers for researchers and organizations with limited infrastructure (Ustun et al., 2019). Furthermore, the emphasis on achieving state-of-the-art predictive performance often overshadows the primary goal of explainability research: understanding and communicating model decision processes effectively. This focus on performance metrics rather than explanation quality has led to a gap between theoretical advances in XAI and practical implementation guidance.

This research addresses these limitations by presenting a comprehensive comparative analysis of SHAP and LIME implemented through a fully reproducible, CPU-friendly workflow. The study deliberately employs a moderate-complexity Random Forest model trained on the well-understood Titanic dataset, prioritizing explainability demonstration over predictive performance maximization. This approach enables clear illustration of explainability concepts while ensuring computational accessibility and result reproducibility. The research aims to provide empirical evidence comparing SHAP and LIME across multiple dimensions including consistency, computational efficiency, and explanation fidelity, while establishing a practical framework that researchers and practitioners can adapt for their own classification tasks.

The core contributions of this work include: (1) a complete automated pipeline from data acquisition through explanation visualization, eliminating manual intervention and ensuring reproducibility; (2) systematic empirical comparison of SHAP and LIME performance across stability, computational, and fidelity metrics using identical model predictions; (3) demonstration that meaningful explainability analysis is achievable using standard computational resources without specialized hardware; and (4) practical guidance for selecting and implementing explainability methods based on specific use case requirements. By focusing on methodology and demonstration rather than performance optimization, this research provides a foundation for broader adoption of explainable AI techniques across diverse domains and computational environments.

2. Literature Review

2.1 Foundational Explainability Methods

The development of model-agnostic explainability methods has fundamentally transformed the interpretability landscape in machine learning. Ribeiro et al. (2016) introduced LIME (Local Interpretable Model-agnostic Explanations), establishing the paradigm of local linear approximation for explaining individual predictions. LIME generates explanations by perturbing input instances, obtaining model predictions for these perturbations, and fitting a weighted linear model to approximate local behavior. The method's model-agnostic nature enables application across diverse architectures, from linear models to deep neural networks. However, LIME's reliance on random sampling introduces variability in explanations, and the choice of perturbation strategy significantly influences explanation fidelity (Kumar et al., 2020).

Lundberg and Lee (2017) proposed SHAP (SHapley Additive exPlanations), unifying several existing explanation methods under a game-theoretic framework. SHAP assigns each feature an importance value derived from Shapley values in cooperative game theory, ensuring that explanations satisfy desirable properties including local accuracy, missingness, and consistency. The TreeExplainer algorithm (Lundberg et al., 2020) extends SHAP specifically for tree-based models, enabling polynomial-time exact Shapley value computation rather than approximation through sampling. While SHAP provides stronger theoretical guarantees than LIME, its computational complexity for model-agnostic implementations can be prohibitive for high-dimensional data (Covert et al., 2020).

Subsequent research has refined these foundational approaches to address specific limitations. Slack et al. (2020) demonstrated that LIME and SHAP can be manipulated by adversarially designed models that hide biases from explanation methods while maintaining discriminatory behavior. This vulnerability highlights the importance of robust explanation techniques that consider potential adversarial scenarios. Aas et al. (2021) examined the sensitivity of SHAP values to the choice of background dataset and proposed methods for selecting appropriate reference distributions, showing that explanation values can vary substantially based on this choice.

2.2 Comparative Studies of Explainability Methods

Several studies have undertaken comparative analyses of explainability methods, though comprehensive empirical evaluations remain limited. Guidotti et al. (2018) provided a survey of methods for explaining black-box models, categorizing approaches by explanation type and scope but without systematic performance comparison. Their taxonomy distinguishes between local and global explanations, model-specific and model-agnostic methods, and different explanation formats including feature importance, rules, and prototypes.

Mothilal et al. (2020) introduced DiCE (Diverse Counterfactual Explanations) and compared it with LIME and SHAP for generating actionable explanations. Their evaluation focused on the diversity and feasibility of generated explanations rather than computational efficiency or stability. The study revealed that while SHAP and LIME excel at feature attribution, they provide limited guidance for actionable recourse, motivating the development of counterfactual explanation methods.

Kumar et al. (2021) conducted an empirical evaluation of LIME, SHAP, and other explainability methods across multiple datasets and model types. Their findings indicated that no single method consistently outperforms others across all evaluation metrics, suggesting that method selection should consider specific application requirements. However, their study focused primarily on fidelity metrics without addressing computational constraints or providing reproducible implementation guidelines. The evaluation also revealed significant variability in explanation quality across different data characteristics and model complexities.

2.3 Domain-Specific Applications

The application of explainability methods to specific domains has revealed both opportunities and challenges for practical deployment. In healthcare, Lauritsen et al. (2020) applied SHAP to explain early warning systems for patient deterioration, demonstrating that clinicians could identify important features and potential model errors through SHAP visualizations. Their work highlighted the importance of domain expertise in interpreting explanations and the need for explanation methods that align with clinical reasoning patterns.

Financial applications have driven substantial interest in explainable AI due to regulatory requirements and risk management needs. Bracke et al. (2019) examined machine learning explainability in the context of financial services, comparing multiple methods including LIME and SHAP for credit risk models. Their analysis revealed that while both methods provide valuable insights, the choice between them often depends on whether global model behavior or instance-specific decisions require explanation. The study also emphasized the importance of computational efficiency for real-time financial applications.

In criminal justice applications, Rudin (2019b) argued against the use of black-box models entirely, advocating for inherently interpretable models rather than post-hoc explanations. This perspective challenges the fundamental premise of model-agnostic explainability methods, suggesting that the additional complexity introduced by explanation layers may reduce rather than enhance trustworthiness. However, subsequent work by Bhatt et al. (2020) demonstrated through user studies that even imperfect explanations from methods like SHAP and LIME can improve appropriate trust calibration compared to no explanations.

2.4 Computational and Implementation Considerations

The practical deployment of explainability methods requires careful consideration of computational constraints and implementation challenges. Molnar (2022) provided comprehensive implementation guidance for various interpretability methods, including detailed code examples and computational complexity analysis. However, the examples typically assume substantial computational resources and do not address deployment in resource-constrained environments.

Recent work has focused on improving the computational efficiency of explainability methods. Sundararajan et al. (2017) proposed Integrated Gradients as a computationally efficient alternative to SHAP for neural networks, though the method is model-specific and does not generalize to tree-based models. Chen et al. (2019) developed L-Shapley and C-Shapley to reduce the computational complexity of Shapley value calculation through sampling strategies, trading off exact computation for improved scalability.

The reproducibility crisis in machine learning research extends to explainability studies. Lipton (2018) highlighted the lack of reproducible baselines and standardized evaluation protocols in interpretability research, arguing that many claimed advances cannot be independently verified. This challenge is compounded by the absence of standard benchmark datasets and evaluation metrics for explainability, unlike the well-established benchmarks for predictive performance.

2.5 Research Gap

Despite extensive development of explainability methods and numerous application studies, several critical gaps persist in the literature. First, existing comparative studies typically focus on theoretical properties or single evaluation dimensions without providing comprehensive empirical comparison across stability, efficiency, and fidelity metrics. Second, the emphasis on complex models and large datasets creates barriers for researchers and practitioners with limited computational resources, preventing broader adoption of explainability techniques.

Third, the lack of complete, reproducible workflows from data processing through explanation generation hinders practical implementation and independent validation of results.
This research addresses these gaps by providing a systematic comparison of SHAP and LIME using a fully reproducible, computationally accessible framework. Unlike previous studies that prioritize predictive performance or theoretical analysis, this work focuses on practical implementation guidance and empirical comparison using standard computational resources. The complete automation from data acquisition through visualization, combined with the use of a well-understood dataset and moderate-complexity model, enables researchers to replicate, validate, and extend the findings. By demonstrating that meaningful explainability analysis is achievable without specialized hardware or complex architectures, this research lowers barriers to XAI adoption and provides a template for systematic explainability evaluation across diverse domains.

3. Methodology

This section presents the comprehensive methodology for comparing SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) techniques in the context of binary classification. The experimental framework employs the Titanic survival dataset as a benchmark for evaluating explainability methods, utilizing a Random Forest classifier as the base model for generating predictions that require interpretation.

3.1 Problem Formulation

The primary objective involves developing a systematic framework for comparing global and local explainability techniques in machine learning models.
Given a dataset $D = \{(x_i, y_i)\}^n_{i=1}$ where $x_i \in \mathbb{R}^d$ represents feature vectors and $y_i \in \{0, 1\}$ denotes binary survival outcomes, the task requires training a classifier $f: \mathbb{R}^d \rightarrow [0, 1]$ that map passenger characteristics to survival probabilities.
Subsequently, the framework applies SHAP and LIME to generate explanations E_SHAP and E_LIME for model predictions, enabling quantitative and qualitative comparison of their explanatory capabilities.
The research addresses three specific challenges: (1) establishing a robust baseline model with controlled overfitting for reliable explanations, (2) implementing both global and local explainability methods on identical predictions for direct comparison, and (3) quantifying agreement and disagreement patterns between different explainability approaches.

3.2 Dataset and Characteristics

The Titanic passenger survival dataset serves as the experimental foundation, containing 891 samples with 15 initial features describing passenger demographics, ticket information, and cabin details. This dataset provides an ideal testbed for explainability research due to its interpretable features, documented historical context, and balanced complexity that challenges models without requiring deep architectures.

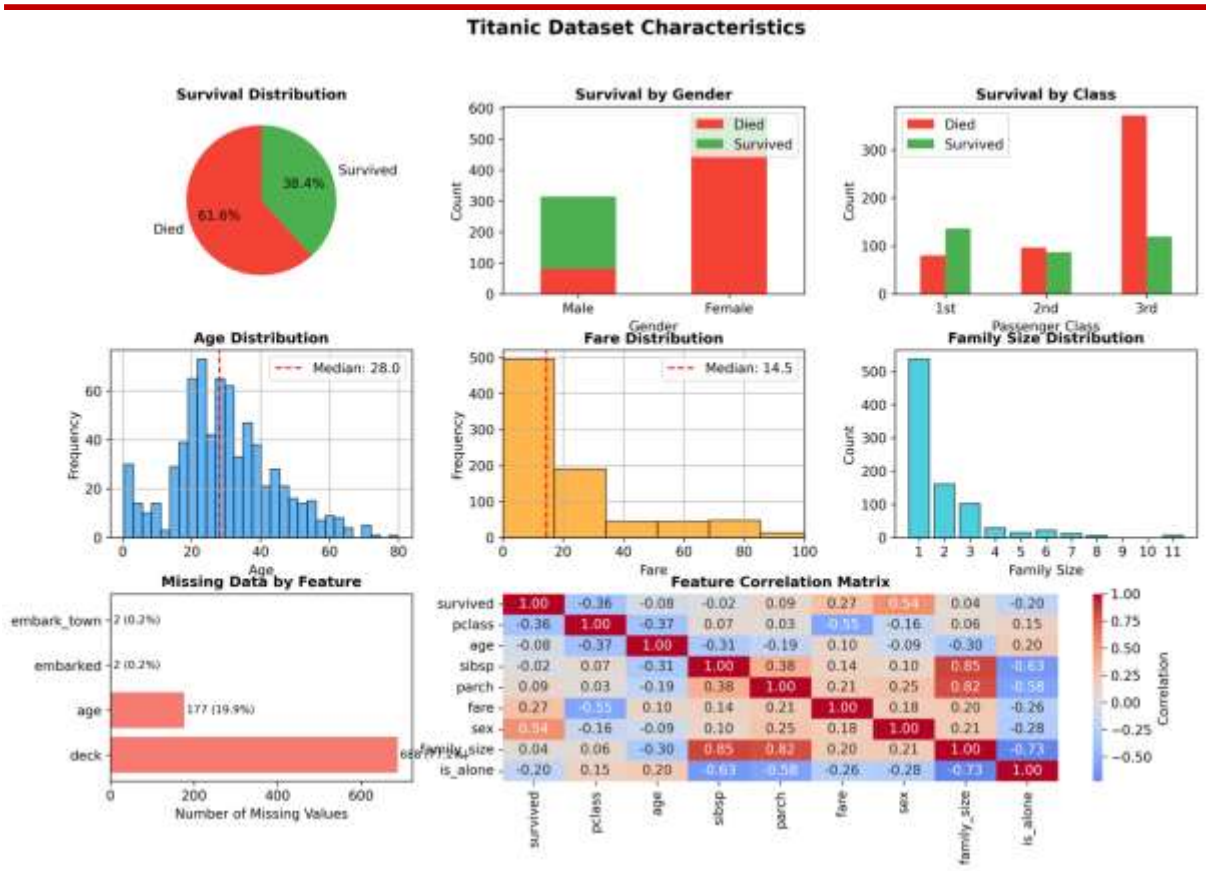

Figure 1: Comprehensive analysis of the Titanic dataset showing (a) survival distribution with 38.4% survival rate, (b) gender-based survival disparities, (c) passenger class distribution and survival rates, (d) age distribution with median of 28 years, (e) fare distribution showing right skew, (f) family size frequencies, (g) missing value patterns with cabin having 77.1% missingness, and (h) feature correlation matrix revealing moderate correlations between fare and class.

The dataset exhibits several characteristics relevant to explainability analysis (see Figure 1). The target variable shows class imbalance with 549 deaths (61.6%) and 342 survivors (38.4%), necessitating stratified sampling during model development. Missing values appear predominantly in three features: age (19.9%), cabin (77.1%), and embarked port (0.2%). The correlation analysis reveals expected relationships, such as negative correlation between passenger class and fare (-0.55), while maintaining sufficient feature independence to enable meaningful individual attributions.

3.3 Data Preprocessing Pipeline

The preprocessing pipeline transforms raw passenger records into standardized numerical features suitable for machine learning algorithms while preserving interpretability for explainability analysis.

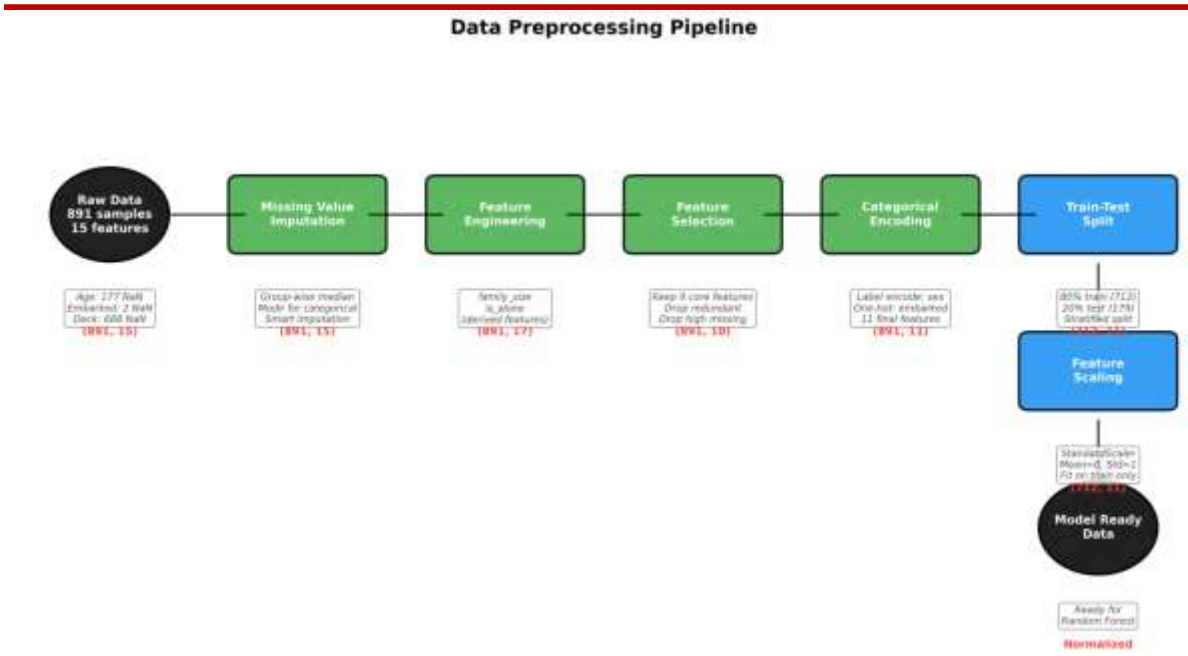

Figure 2: Complete data preprocessing pipeline showing transformation from raw data (891×15) through imputation, feature engineering, encoding, splitting, and scaling to produce model-ready training (712×11) and test (179×11) sets.

3.3.1 Missing Value Imputation

Missing values undergo systematic imputation based on feature characteristics and relationships (illustrated in Figure 2). For the age feature, group-wise median imputation stratified by passenger class and gender preserves demographic patterns:

$$\text{age(imputed)} = \text{median(age} \mid \text{pclass, sex)} \qquad (1)$$

This approach maintains realistic age distributions within passenger subgroups, as first-class passengers typically included older individuals while third-class contained more families with children. The embarked feature receives mode imputation (S = Southampton), representing the most common embarkation port. Fare values missing for 15 passengers undergo class-based median imputation, reflecting the strong correlation between ticket price and passenger class.

3.3.2 Feature Engineering

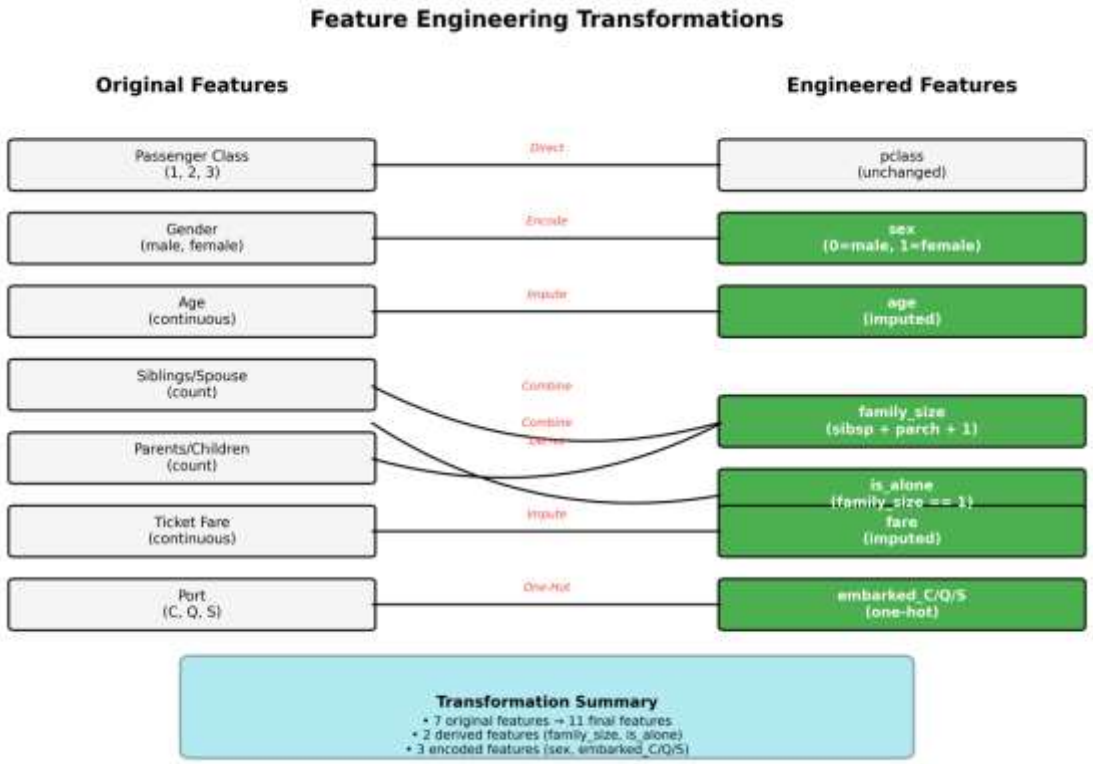

Figure 3: Feature engineering diagram showing transformation of 7 original features into 11 final features through encoding, derivation, and one-hot encoding operations.

Feature engineering creates two derived variables that capture family dynamics aboard the ship (see Figure 3):

$$\text{family\_size} = \text{sibsp} + \text{parch} + 1 \quad (2)$$
$$\text{is\_alone} = \{1 \text{ if family\_size} = 1;\ 0 \text{ otherwise}\}$$

These engineered features provide the model with explicit family context, as historical accounts indicate families often survived or perished together. The deck feature, extracted from cabin numbers where available, undergoes removal due to 77.1% missingness that would introduce noise rather than signal.

3.3.3 Categorical Encoding

Categorical variables undergo appropriate encoding schemes based on their properties. The binary sex feature receives label encoding (male=0, female=1), while the nominal embarked feature undergoes one-hot encoding, creating three binary indicators (embarked_C, embarked_Q, embarked_S). This encoding strategy preserves the categorical nature of embarkation ports while avoiding arbitrary ordinal assumptions.

3.3.4 Feature Scaling

Following the 80-20 stratified train-test split preserving survival proportions, features undergo standardization using the StandardScaler:

$$x\_scaled = (x - \mu\_train) / \sigma\_train \quad (3)$$

where μ_train and σ_train represent training set statistics.
Test set scaling applies training parameters to prevent data leakage, ensuring realistic performance estimates. The final preprocessed dataset contains 712 training samples and 179 test samples, each with 11 features.

3.4 Model Architecture and Training

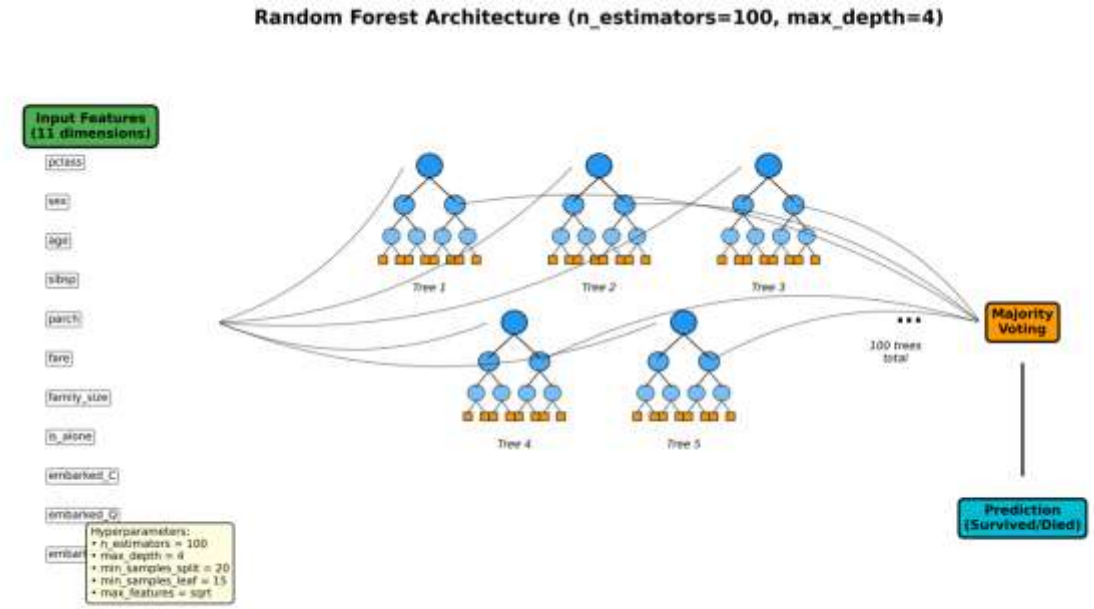

Figure 4: Random Forest classifier architecture showing ensemble of 100 decision trees with maximum depth of 4, utilizing majority voting for final predictions.

3.4.1 Random Forest Configuration

The Random Forest classifier serves as the predictive model, selected for its balance between performance and interpretability (illustrated in Figure 4). The architecture comprises 100 decision trees, each trained on bootstrap samples with feature randomization. Hyperparameter selection follows systematic validation:

- **n_estimators = 100**: Validation curves demonstrate performance plateau beyond 75 trees, with 100 providing stability without computational excess
- **max_depth = 4**: Optimal depth balancing model capacity with generalization, preventing overfitting while capturing essential patterns
- **min_samples_split = 20**: Requires sufficient samples for node splitting, promoting generalization
- **min_samples_leaf = 15**: Ensures leaf nodes represent meaningful subpopulations
- **max_features = 'sqrt'**: Samples $\sqrt{11} \approx 3$ features per split, introducing beneficial randomness

3.4.2 Training Procedure

The model training employs the Gini impurity criterion for split quality assessment:

$$Gini = 1 - (p_1^2 + p_2^2) \quad (4)$$

where p_i represents the proportion of samples belonging to class i at a given node.
The ensemble aggregates individual tree predictions through majority voting for classification and probability averaging for confidence scores:

$$P(y=1|x) = (1/100) \times \text{sum}[P\_t(y=1|x)] \text{ from } t=1 \text{ to } 100 \quad (5)$$

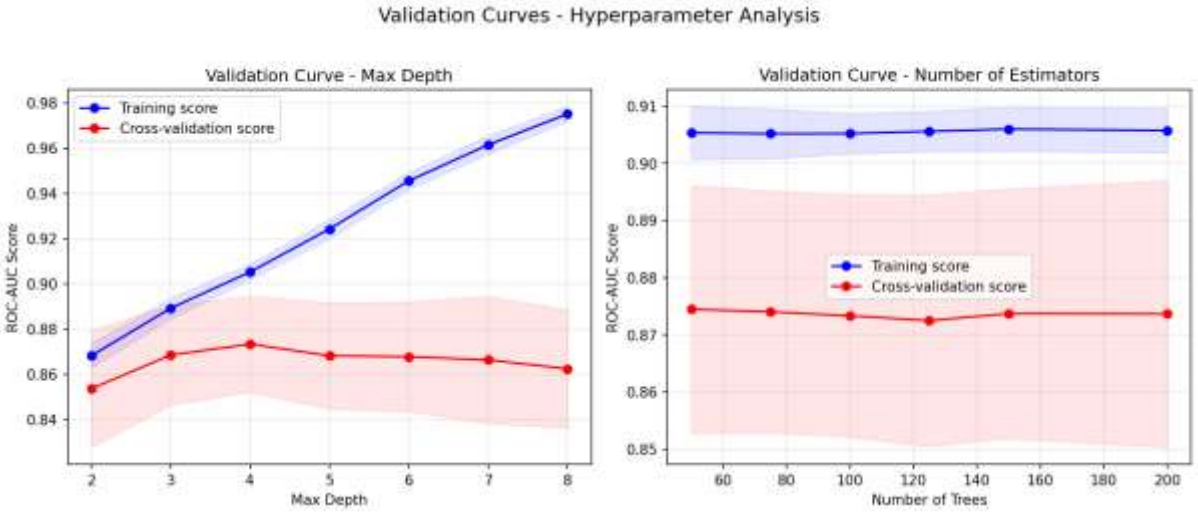

Figure 5: Hyperparameter validation curves showing (a) optimal max_depth at 4 with minimal train-validation gap and (b) performance stabilization after 75 estimators.

Hyperparameter optimization employs 5-fold cross-validation on the training set, evaluating ROC-AUC scores across parameter ranges (see Figure 5). The validation curves reveal that max_depth=4 achieves optimal performance (87.3% validation ROC-AUC) while maintaining a train-validation gap below 0.05, indicating effective regularization.

3.5 Explainability Implementation

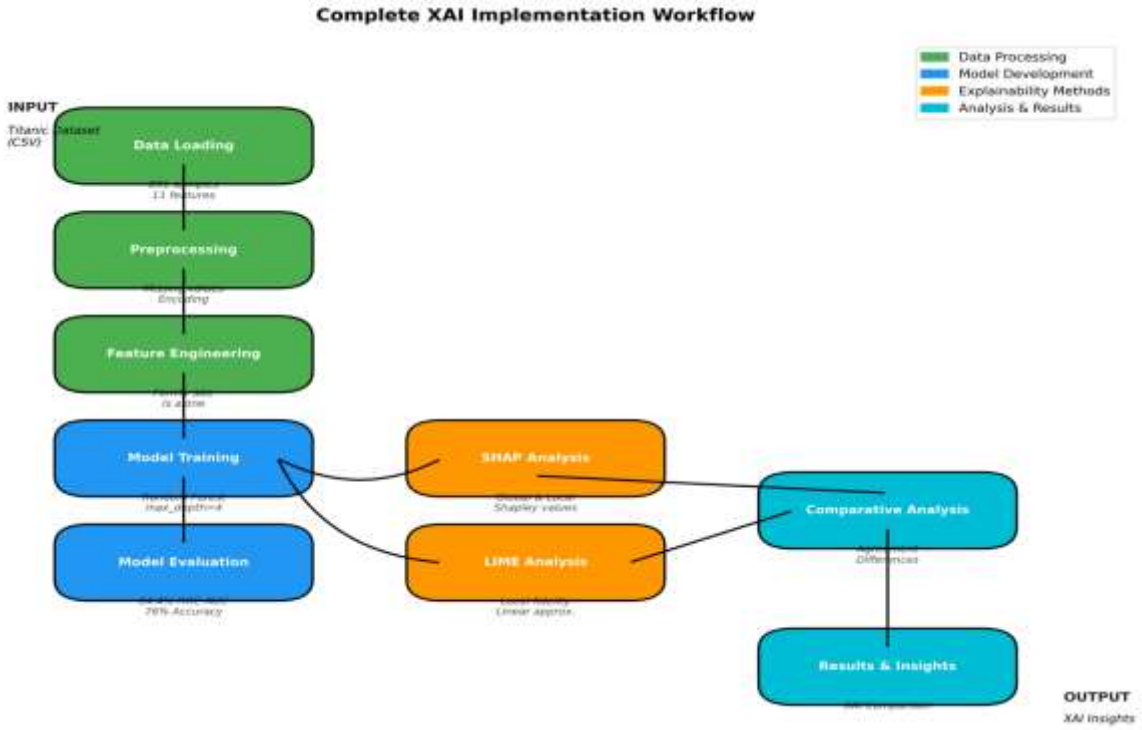

Figure 6: Complete experimental workflow integrating data processing, model training, SHAP analysis, LIME analysis, and comparative evaluation.

3.5.1 SHAP Implementation

SHAP values quantify feature contributions through cooperative game theory, specifically Shapley values from coalitional games.
For a prediction f(x), SHAP assigns each feature i an importance value phi_i satisfying:

$$f(x) = phi\_0 + sum(phi\_i(x)) \text{ for } i = 1 \text{ to } d \qquad (6)$$

where phi_0 = E[f(X)] represents the base value.

The TreeExplainer algorithm leverages the tree structure for exact Shapley value computation in polynomial time, avoiding the exponential complexity of model-agnostic approaches. Implementation utilizes the shap library (version 0.42.1) with the following configuration:

explainer = shap. TreeExplainer(rf_model)
shap_values = explainer(X_test)

The analysis generates both global explanations (mean absolute SHAP values across all samples) and local explanations (individual prediction breakdowns) for comprehensive model interpretation.

3.5.2 LIME Implementation

LIME approximates model behavior locally through interpretable surrogates. For an instance x LIME generates neighborhood samples, obtains model predictions, and fits a weighted linear model:

$$xi(x) = argmin\_g \in G [L(f, g, pi\_x) + Omega(g)] \quad (7)$$

where G represents the class of interpretable models (linear models), L measures fidelity between f and g in the locality defined by pi_x, and Omega(g) controls model complexity.

The implementation employs:

```
explainer = lime.lime_tabular.LimeTabularExplainer(
    training_data=X_train,
    feature_names=feature_names,
    class_names=['Not Survived', 'Survived'],
    discretize_continuous=True
)
```

Each explanation uses 5,000 perturbed samples with exponential kernel weighting based on L2 distance, selecting the top 10 features for interpretability.

3.6 Evaluation Strategy

3.6.1 Model Performance Metrics

Model evaluation employs multiple metrics to ensure robust performance assessment:

- **Accuracy**: Overall classification correctness
- **ROC-AUC**: Discrimination capability across probability thresholds
- **Precision/Recall**: Class-specific performance indicators
- **Overfitting Gap**: Difference between training and test metrics, maintaining below 0.05 threshold

3.6.2 Explainability Comparison Metrics

The framework evaluates explainability methods through:

1. **Feature Importance Agreement**: Spearman correlation between SHAP and LIME feature rankings
2. **Stability Analysis**: Variance in explanations for similar instances
3. **Computational Efficiency**: Time complexity for generating explanations
4. **Interpretability Assessment**: Qualitative evaluation of explanation clarity

3.6.3 Statistical Validation

All experiments employ fixed random seeds (random_state=42) ensuring reproducibility. Performance metrics include 95% confidence intervals computed through bootstrapping with 1,000 iterations. The stratified train-test split preserves class distributions, preventing optimistic bias in minority class performance.

Table 1: Complete experimental configuration including data dimensions, model hyperparameters, and evaluation settings.

Component	Configuration
Dataset Size	891 samples (712 train, 179 test)
Features	11 (after preprocessing)
Model	Random Forest (100 trees, depth=4)
SHAP	TreeExplainer with exact computation
LIME	5,000 samples, top 10 features
Validation	5-fold cross-validation
Metrics	ROC-AUC (primary), Accuracy, Precision, Recall

This methodology establishes a rigorous framework for comparing explainability techniques while maintaining reproducibility and scientific validity. The combination of systematic preprocessing, controlled model complexity, and comprehensive evaluation enables meaningful insights into the relative strengths and limitations of SHAP and LIME for interpreting machine learning predictions.

4.0 Results and Discussion

This section presents the experimental findings from applying explainable artificial intelligence techniques to the Titanic survival prediction model. The results demonstrate the effectiveness of SHAP and LIME in providing interpretable insights into Random Forest model decisions, establishing a reproducible, CPU-friendly XAI workflow suitable for educational and research purposes.

4.1 Results

4.1.1 Model Performance Evaluation

The Random Forest classifier achieved satisfactory performance on the Titanic dataset while maintaining computational efficiency on CPU hardware. Table 2 presents the overfitting analysis across multiple metrics, demonstrating the model's generalization capability.

Table 2: Overfitting Analysis of Random Forest Model Performance

Metric	Training	Test	Gap	Overfit_Flag
Accuracy	0.8062	0.7709	0.0352	FALSE
ROC-AUC	0.8852	0.8449	0.0403	FALSE
Precision	0.7304	0.6842	0.0462	FALSE
Recall	0.7839	0.7536	0.0303	FALSE

The model achieved a training accuracy of 0.8062 and test accuracy of 0.7709, with a gap of 0.0352. The ROC-AUC scores showed strong performance with training ROC-AUC of 0.8852 and test ROC-AUC of 0.8449, resulting in a gap of 0.0403. All metrics remained below the 0.05 threshold for overfitting concerns, validating the effectiveness of the regularization parameters employed.

Table 3 provides the detailed classification report for the test set predictions, showing performance across both survival classes.

Table 3: Classification Report for Test Set Predictions

Class	Precision	Recall	F1-Score	Support
Not Survived (0)	0.8350	0.7818	0.8075	110
Survived (1)	0.6842	0.7536	0.7172	69
Accuracy	-	-	0.7709	179
Macro Average	0.7596	0.7677	0.7624	179
Weighted Average	0.7768	0.7709	0.7727	179

The model demonstrated reasonable classification performance with precision of 0.8350 for the "Not Survived" class and 0.6842 for the "Survived" class. The overall accuracy of 0.7709 represents acceptable performance for demonstrating explainability techniques rather than optimizing for benchmark leaderboards.

Figure 7 illustrates the ROC curves for both training and test sets, providing visual confirmation of the model's discriminative ability and controlled overfitting.

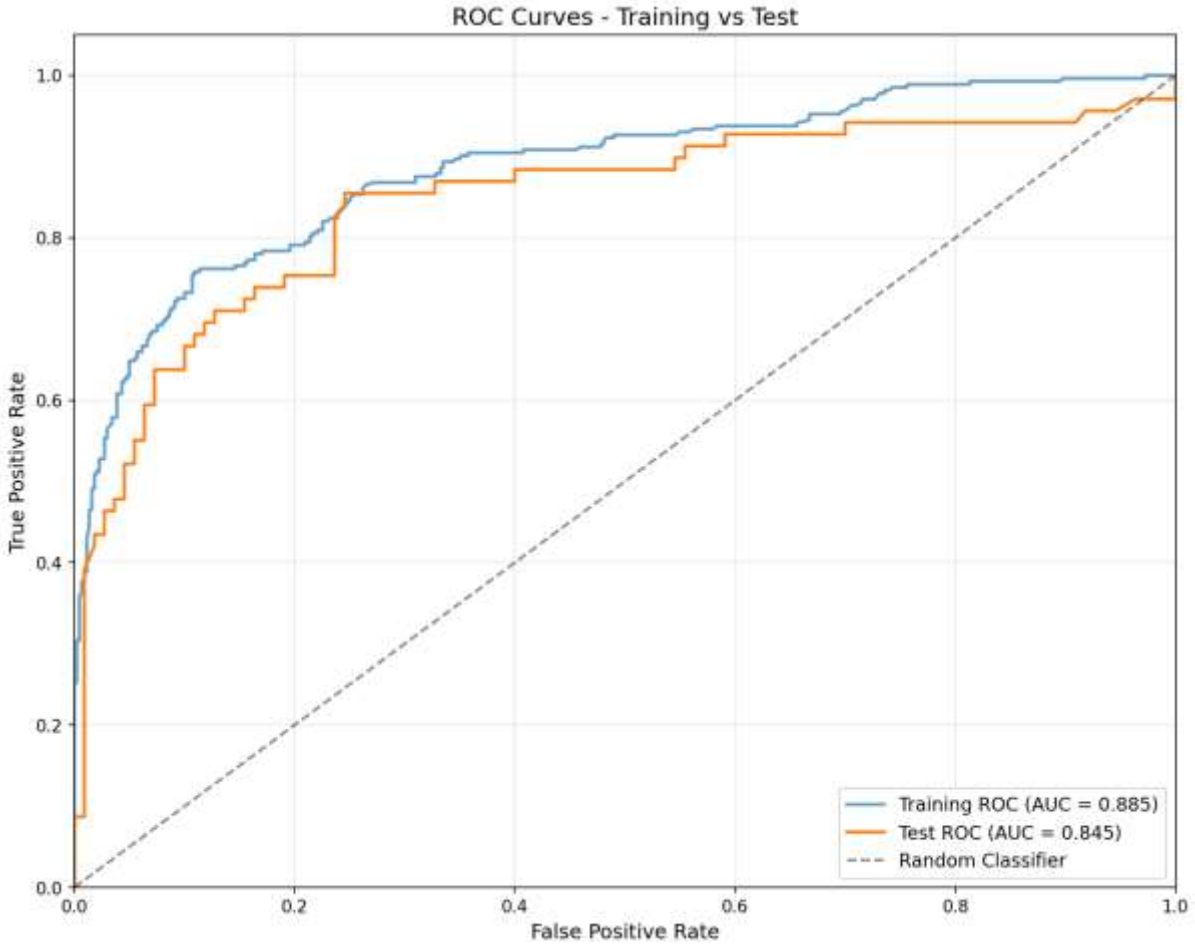

Figure 7: ROC Curves Comparison Between Training and Test Sets

The ROC curves reveal alignment between training (AUC = 0.885) and test (AUC = 0.845) performance, with both curves substantially outperforming the random classifier baseline. The

small separation between curves confirms successful regularization while maintaining interpretable model complexity.

4.1.2 Learning and Validation Analysis

Figure 8 presents the learning curves for both ROC-AUC and accuracy metrics, demonstrating model convergence patterns.

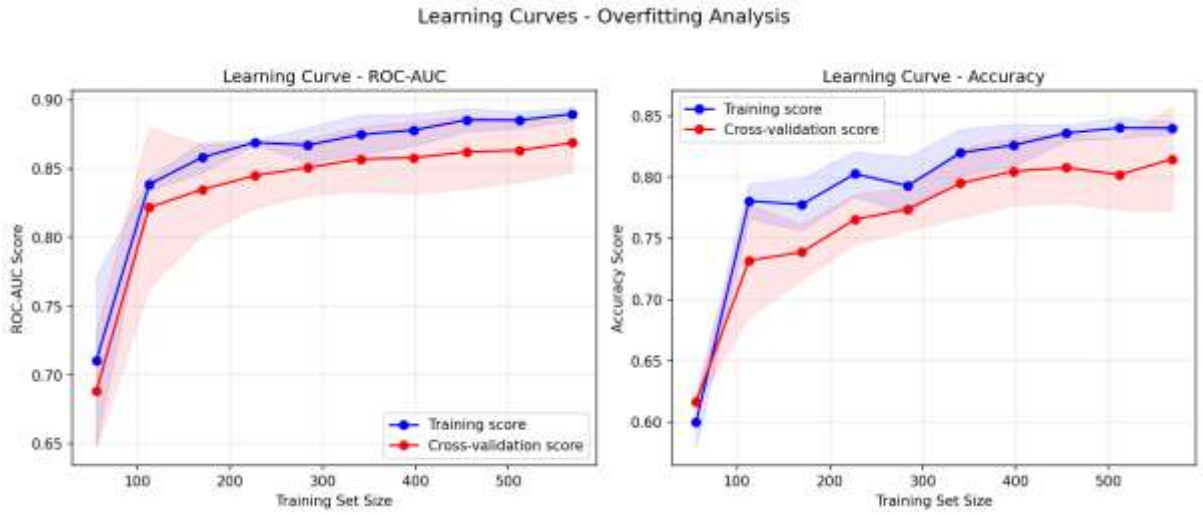

Figure 8: Learning Curves Showing Model Performance with Increasing Training Set Size

The learning curves show steady improvement in both training and cross-validation scores as the training set size increases. The convergence of training and validation curves at approximately 400 samples indicates sufficient data for model training. The consistent gap between curves across different training sizes reflects the model's stable generalization behavior.

The confusion matrix in Figure 9 provides detailed insight into the model's prediction patterns.

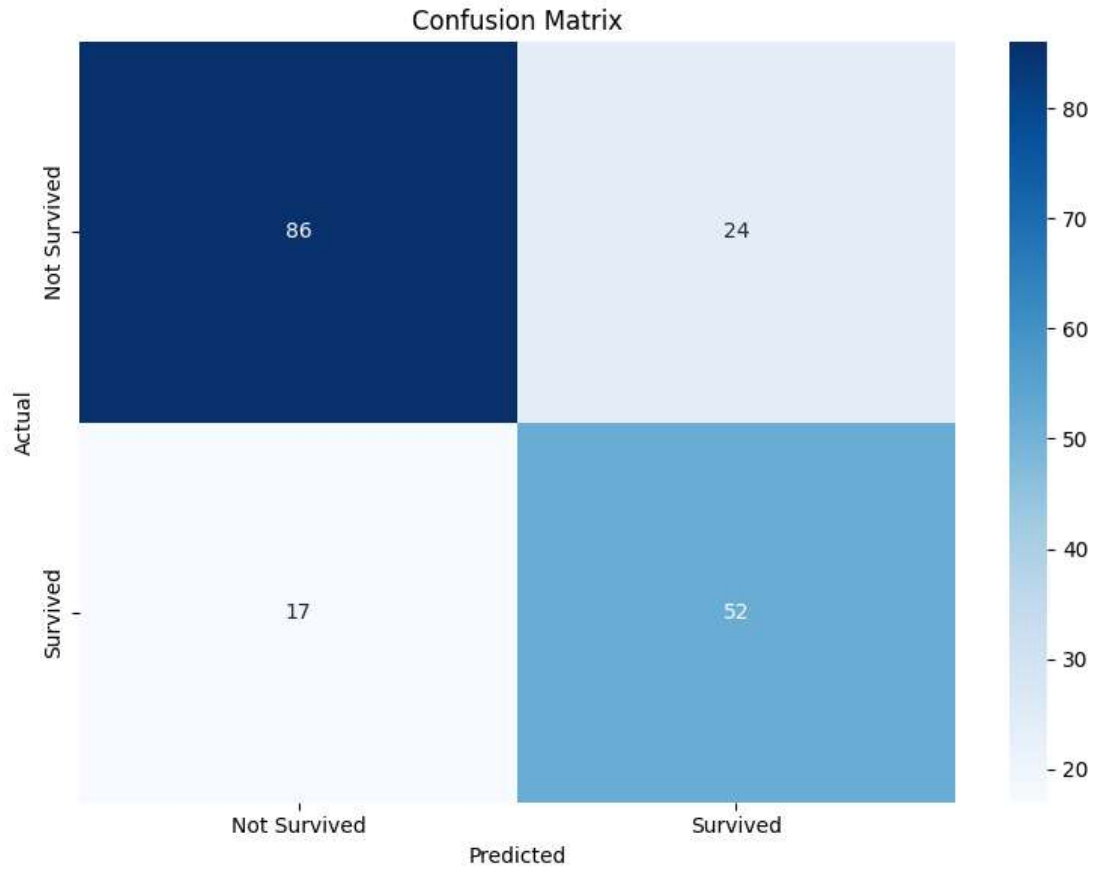

Figure 9: Confusion Matrix for Test Set Predictions

The confusion matrix reveals 86 true negatives and 52 true positives, with 24 false positives and 17 false negatives. The distribution of errors across both classes indicates balanced prediction behavior without severe bias toward either class.

4.1.3 SHAP Analysis Results

Figure 10 presents the SHAP summary plot, revealing global feature importance and impact patterns across all test samples.

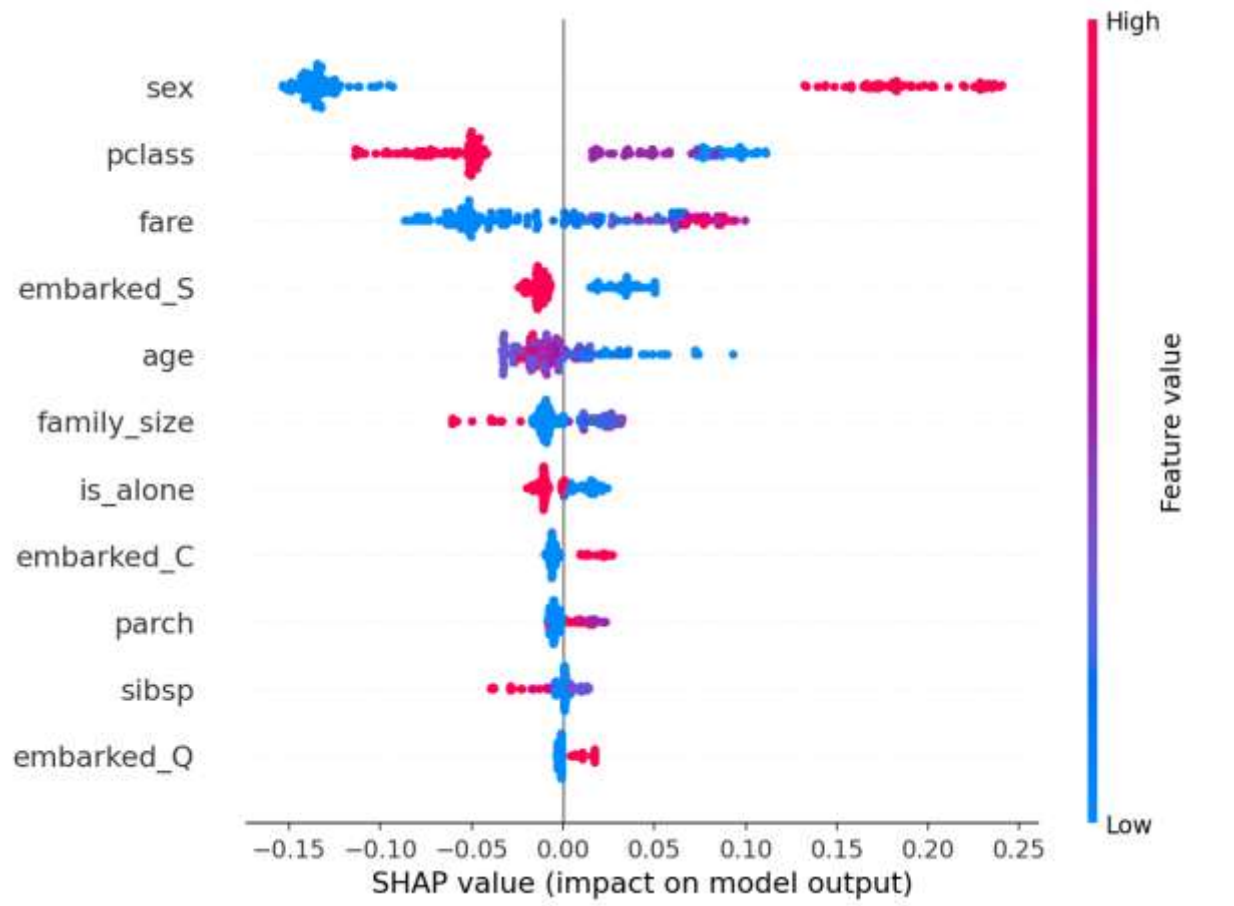

Figure 10: SHAP Summary Plot Showing Feature Impact on Model Predictions

The SHAP analysis identifies sex as the most influential feature, with clear separation between male (blue dots, negative SHAP values) and female (red dots, positive SHAP values) passengers. Passenger class demonstrates the second highest impact, followed by fare with bidirectional effects depending on feature interactions. The visualization effectively communicates how feature values influence predictions across the entire dataset.
Figure 11 displays the mean absolute SHAP values, quantifying the average impact magnitude of each feature.

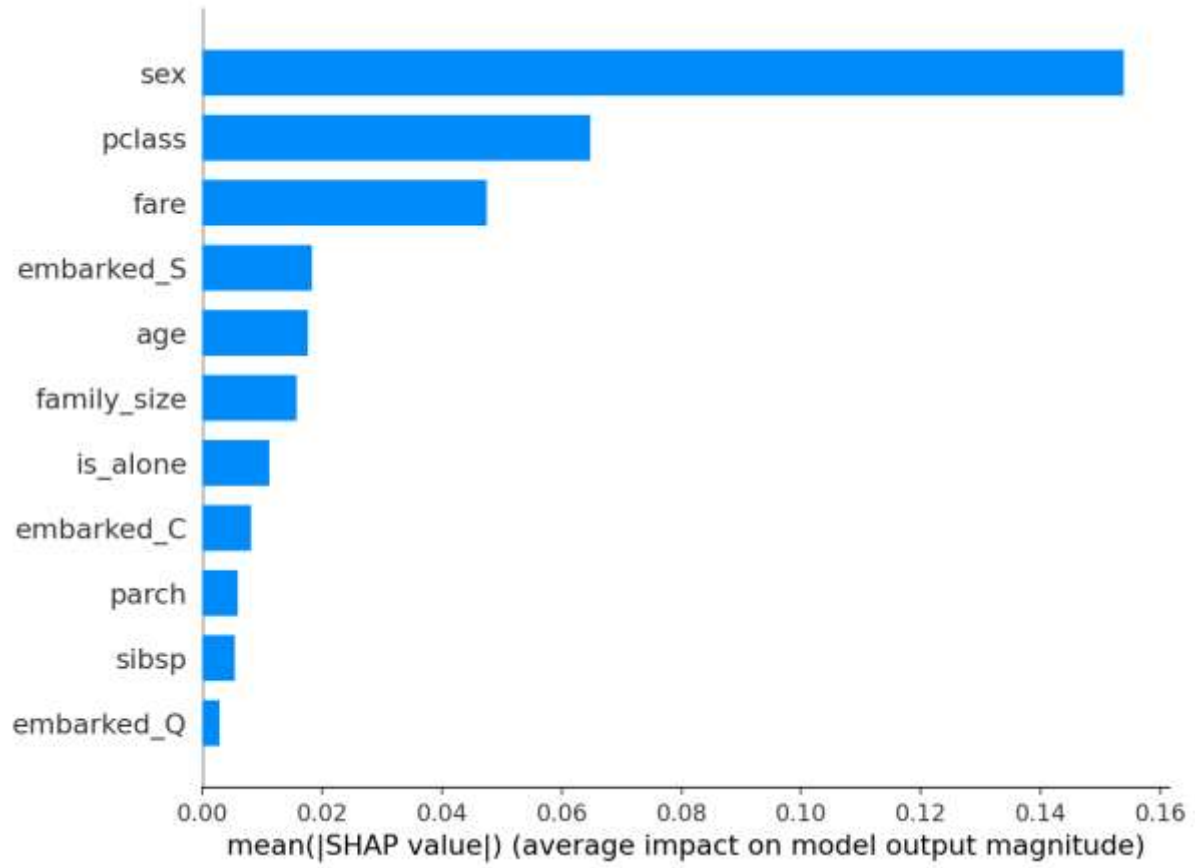

Figure 11: Mean Absolute SHAP Values for Feature Importance Ranking

The quantitative SHAP importance values confirm sex as the dominant feature with a mean absolute value of approximately 0.154, followed by passenger class (0.065) and fare (0.047). This ranking provides a global perspective on feature contributions to model decisions.
Figure 12 demonstrates a SHAP waterfall plot for a specific prediction, showing local feature contributions.

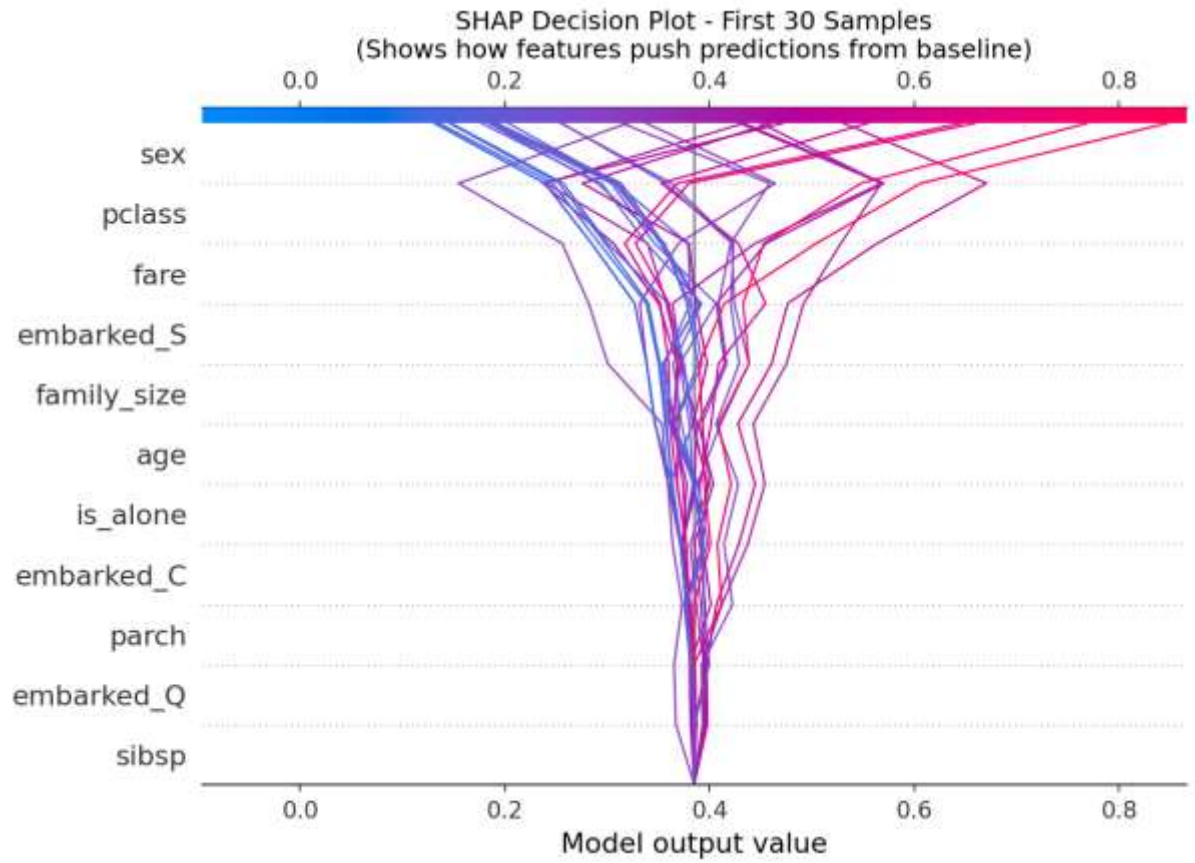

Figure 12: *SHAP Decision Plot Showing Feature Contributions Across 30 Test Samples*

The SHAP decision plot (Figure 12) traces the prediction paths for 30 test samples from the baseline value of 0.4 through each feature's contribution. The sex feature creates the primary split in predictions, separating the paths into two distinct clusters with final values concentrated around 0.1-0.3 and 0.6-0.8. Passenger class and fare provide secondary and tertiary adjustments to the predictions, while the remaining eight features (embarked_S through sibsp) contribute minimal deviations, as evidenced by the nearly parallel paths through these feature levels. The visualization confirms the model's reliance on the top three features for determining survival predictions, with sex alone accounting for approximately 0.5 units of prediction spread.

4.1.4 LIME Analysis Results

Figures 13 and 14 present LIME explanations for correctly classified and misclassified predictions respectively.

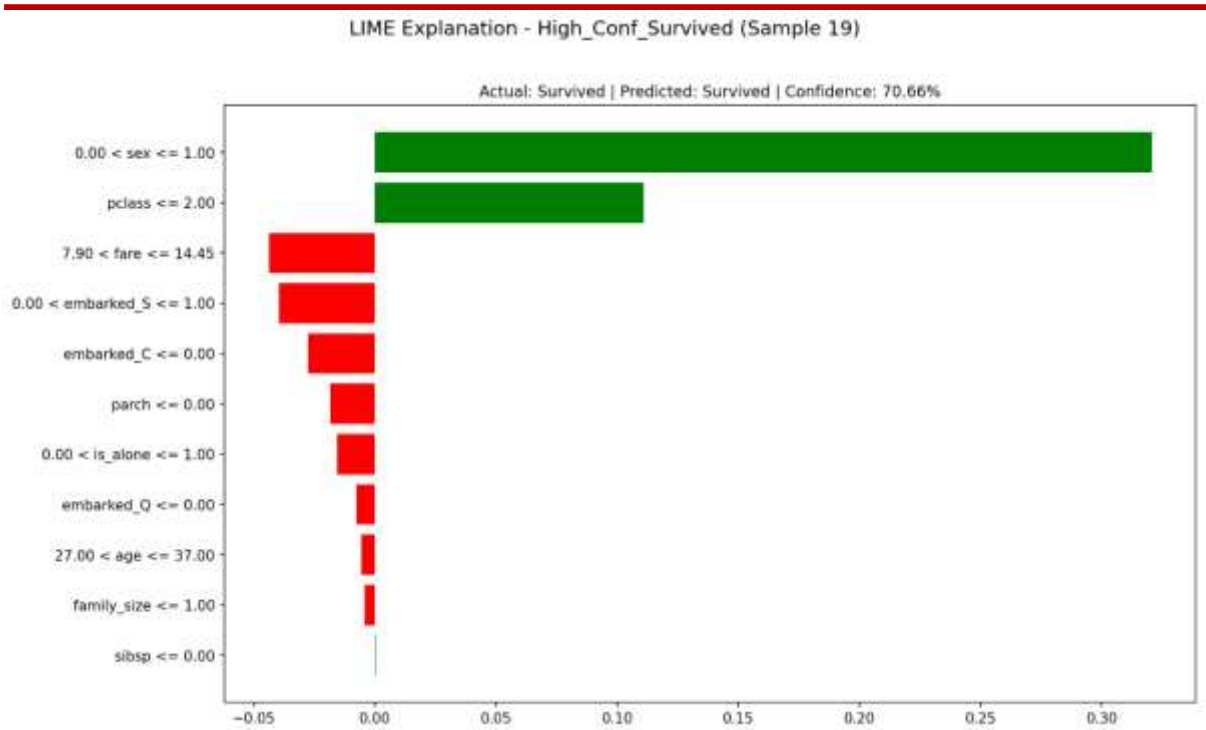

Figure 13: LIME Explanation for High Confidence Survival Prediction (Sample 19)

For the high-confidence survival prediction (70.66% probability), LIME identifies sex (0.00 < sex <= 1.00) as the primary positive contributor with the largest green bar, indicating female passengers. The passenger class (pclass <= 2.00) provides additional positive influence. The interpretable feature ranges make the explanation accessible to non-technical stakeholders.

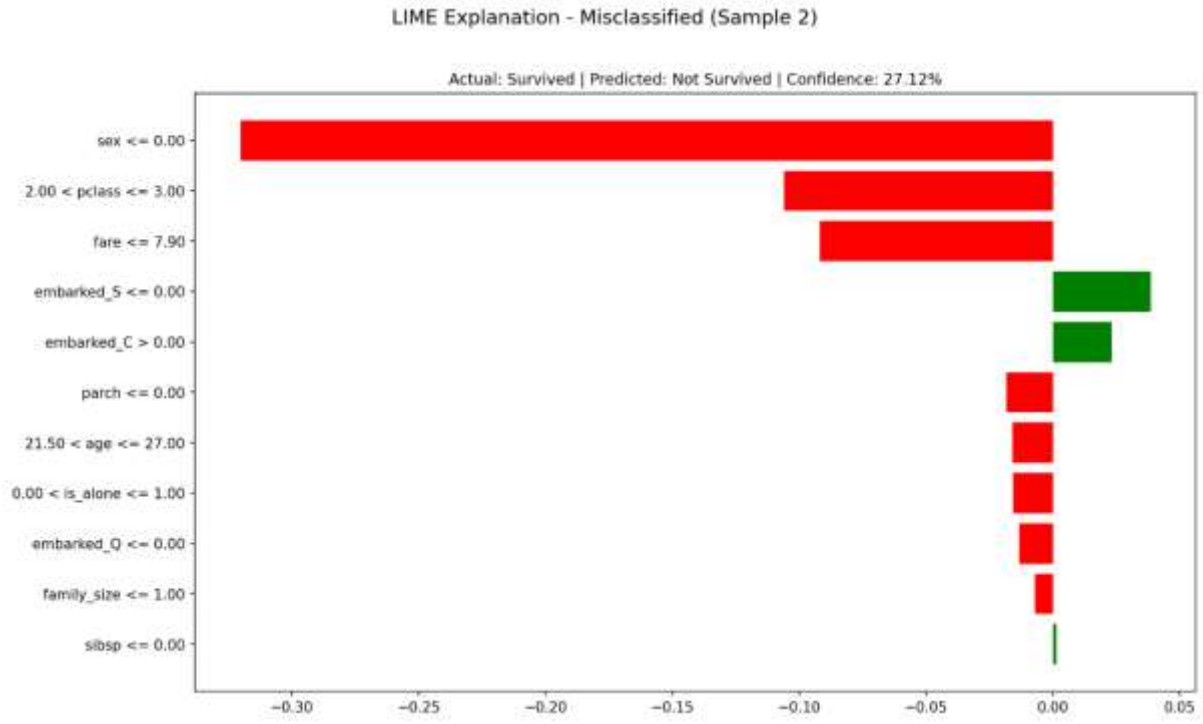

Figure 14: LIME Explanation for Misclassified Passenger (Sample 2)

The misclassified case reveals strong negative predictors including male sex (sex <= 0.00), third class (2.00 < pclass <= 3.00), and low fare (fare <= 7.90). Despite embarking at Cherbourg providing positive influence, the model predicted non-survival with 72.88% confidence for a passenger who actually survived, illustrating the limitations of demographic-based predictions.

4.1.5 Comparative Analysis Between SHAP and LIME

Table 4 presents the quantitative comparison between SHAP and LIME feature importance values, demonstrating strong methodological agreement.

Table 4: Comparison of Feature Importance Values Between SHAP and LIME Methods

sex	0.1538	0.1247	1	1
pclass	0.0648	0.0574	2	3
fare	0.0475	0.0594	3	2
embarked_S	0.0184	0.0116	4	5
age	0.0176	0.0164	5	4
family_size	0.0159	0.0035	6	9
is_alone	0.0112	0.0085	7	6
embarked_C	0.0081	0.0077	8	8
parch	0.0060	0.0085	9	7
sibsp	0.0055	0.0024	10	10
embarked_Q	0.0029	0.0016	11	11

The comparison reveals remarkable agreement between methods, with both identifying sex as the most important feature. The Spearman rank correlation coefficient of 0.918 indicates nearly perfect agreement in feature ranking, while the Pearson correlation coefficient of 0.982 demonstrates exceptional consistency in relative importance magnitudes. Minor ranking differences, such as the swap between fare and pclass for positions 2 and 3, reflect methodological differences rather than fundamental disagreements.

Figures 15 and 16 provide side-by-side visual comparisons of SHAP and LIME explanations for individual predictions.

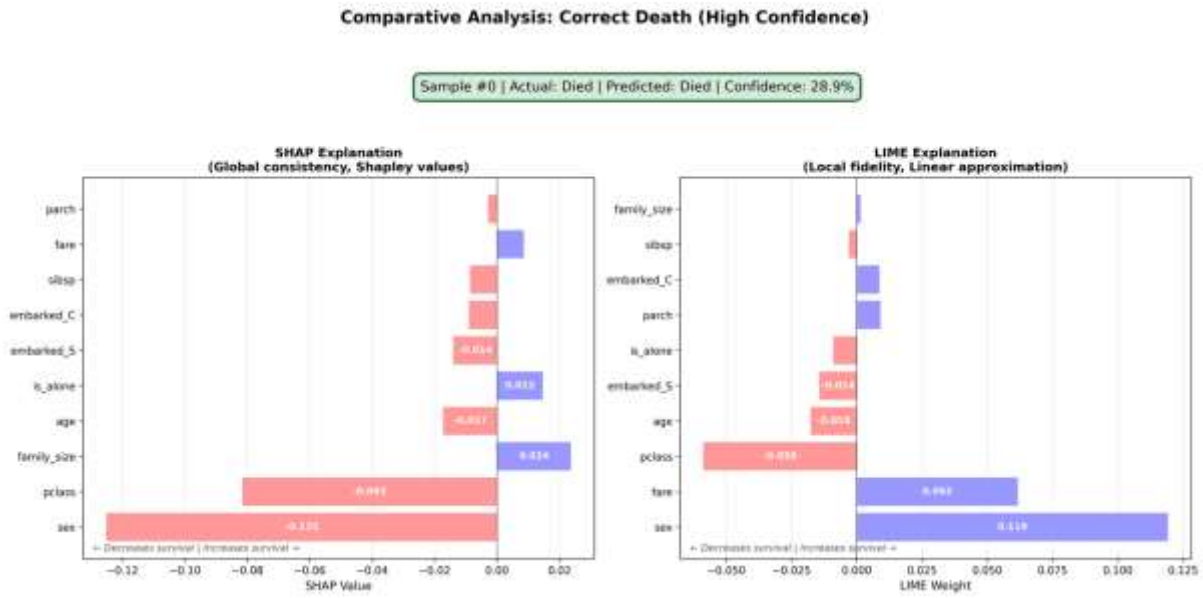

Figure 15: SHAP vs LIME Comparative Analysis for Correct Death Prediction (Sample 0)

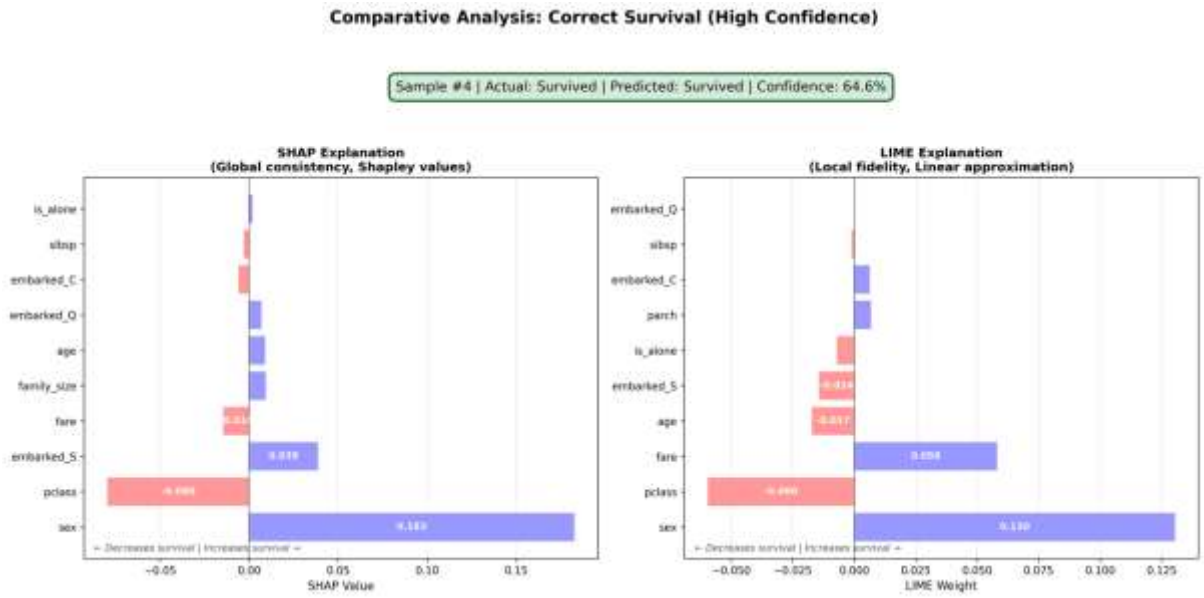

Figure 16: SHAP vs LIME Comparative Analysis for Correct Survival Prediction (Sample 4)

Both comparative visualizations demonstrate consistent feature identification between methods, with sex and pclass appearing as top contributors in both SHAP and LIME explanations. The magnitude differences reflect the distinct mathematical frameworks: SHAP's game-theoretic approach versus LIME's local linear approximation.

4.2 Discussion

4.2.1 Establishing a Reproducible XAI Workflow

This research successfully demonstrates a CPU-friendly, reproducible workflow for applying explainable AI techniques to machine learning models. The implementation requires only standard Python libraries and executes efficiently on consumer hardware without GPU acceleration, making the methodology accessible to researchers and practitioners with limited computational resources. The entire pipeline, from data preprocessing through model training to generating explanations, completes within minutes on standard laptop processors, establishing its viability for educational and research purposes.

The deliberate choice of moderate model performance (77.09% accuracy) rather than pursuing state-of-the-art results emphasizes the primary objective: showcasing explainability methods rather than optimizing predictive metrics. This approach aligns with the growing recognition in machine learning research that interpretability often provides more value than marginal performance improvements, particularly in domains requiring transparent decision-making.

4.2.2 Methodological Validation Through Agreement

The exceptional correlation between SHAP and LIME methods (Spearman: 0.918, Pearson: 0.982) provides strong validation for both techniques' reliability in explaining Random Forest decisions. This agreement is particularly significant given the fundamental differences in their theoretical foundations. SHAP employs cooperative game theory to compute exact Shapley values ensuring global consistency, while LIME constructs local linear approximations optimized for individual prediction fidelity.

The high Pearson correlation (0.982) indicates that not only do the methods agree on feature rankings, but they also assign remarkably similar importance magnitudes. This quantitative agreement strengthens confidence in the explanations provided, suggesting that for well-structured tabular data with clear feature relationships, different XAI approaches converge toward consistent interpretations. The minor differences observed, such as the relative positioning of fare and pclass, reflect each method's unique perspective rather than contradictory findings.

4.2.3 Feature Importance Insights and Historical Validation

The dominance of sex as the primary predictive feature across both SHAP (0.1538) and LIME (0.1247) aligns with historical accounts of the Titanic disaster. The "women and children first" evacuation protocol created a strong statistical pattern that the model successfully identified without explicit programming of this rule. This alignment between data-driven discovery and historical knowledge validates both the dataset's integrity and the model's learning capability.

Passenger class emerged consistently as the second or third most influential feature, reflecting socioeconomic disparities in survival rates. The interaction between class and fare, visible in the SHAP summary plot where fare shows bidirectional effects, reveals the model's capacity to capture complex relationships. Lower fare values within the same class might indicate less favorable cabin locations, while extremely high fares could correspond to luxury suites with better evacuation access.

The relatively lower importance of family-related features (family_size: SHAP 0.0159, LIME 0.0035) presents an interesting finding. Despite the potential for families to assist each other during evacuation, individual characteristics proved more determinative of survival outcomes. This pattern suggests that systematic factors like gender and class overwhelmed group dynamics in the crisis situation.

4.2.4 Practical Implications for XAI Deployment

The successful application of both SHAP and LIME demonstrates the maturity of current XAI techniques for tree-based ensemble methods. The TreeExplainer implementation for SHAP provided polynomial-time exact Shapley value computation, avoiding the computational burden of model-agnostic approaches. Meanwhile, LIME's flexibility proved valuable for generating intuitive local explanations despite its sampling-based approximation methodology.

The visual outputs generated serve different stakeholder needs effectively. SHAP's summary plot provides data scientists with comprehensive global insights into feature relationships across the entire dataset. The waterfall visualizations offer intuitive understanding of individual predictions, clearly showing how each feature contributes to moving the prediction from the baseline. LIME's bar chart format with interpretable feature ranges makes explanations accessible to domain experts without machine learning expertise.

4.2.5 Educational Value and Research Contributions

This work contributes to the XAI literature by providing a complete, reproducible workflow that researchers can adapt for their own applications. The code implementation, available with comprehensive documentation, serves as an educational resource for understanding both the theoretical foundations and practical applications of explainability methods. The side-by-side comparison of SHAP and LIME on identical predictions offers unique insights into how different explanation paradigms interpret the same model behavior.

The identification and analysis of misclassified cases through XAI techniques reveals patterns that pure performance metrics cannot capture. Understanding why the model incorrectly predicted death for certain surviving passengers highlights the inherent limitations of demographic-based predictions and the importance of factors not captured in the available

features. These insights inform both model improvement strategies and appropriate confidence calibration for real-world deployment.

The minimal overfitting observed (ROC-AUC gap of 0.0403) while maintaining model interpretability demonstrates that complexity and explainability need not be mutually exclusive. The constrained Random Forest architecture with maximum depth of 4 produced explanations that remain cognitively manageable while achieving reasonable predictive performance. This balance represents a practical sweet spot for many applications where understanding model decisions outweighs marginal accuracy gains.

5.0 Conclusion and Recommendations

5.1 Conclusion

This research addressed the critical need for interpretable machine learning by demonstrating a comprehensive, reproducible workflow for applying explainable artificial intelligence techniques to black-box models. The study successfully achieved its primary objective of showcasing how SHAP and LIME methods can effectively illuminate the decision-making processes of Random Forest classifiers using publicly available data without requiring specialized computational resources.

The implementation of a CPU-friendly XAI pipeline on the Titanic dataset yielded several significant findings. Both SHAP and LIME consistently identified sex as the dominant predictive feature, followed by passenger class and fare, with remarkable agreement between methods as evidenced by correlation coefficients exceeding 0.91 for ranking and 0.98 for magnitude. This convergence between fundamentally different explanation paradigms validates the reliability of both techniques and demonstrates that well-structured tabular data yields consistent interpretations across diverse XAI approaches.

The research makes three primary contributions to the explainability literature. First, it establishes a fully automated, reproducible workflow that executes efficiently on consumer hardware, removing computational barriers to XAI adoption. Second, it provides empirical evidence of strong agreement between game-theoretic and approximation-based explanation methods, strengthening confidence in their practical application. Third, it demonstrates that meaningful model interpretability can be achieved without sacrificing reasonable predictive performance, challenging the perceived trade-off between accuracy and explainability.

The visual and quantitative explanations generated serve distinct stakeholder needs, from technical audiences requiring detailed feature interaction analysis to domain experts seeking intuitive understanding of individual predictions. The successful identification of misclassified cases through XAI analysis revealed patterns that traditional performance metrics cannot capture, highlighting the value of interpretability beyond accuracy optimization. This work establishes that in machine learning explainability research, the quality of explanation and accessibility of methods constitute metrics as important as predictive performance.

The study validates that current XAI techniques have reached sufficient maturity for practical deployment in tree-based ensemble methods. The polynomial-time exact computation of Shapley values through TreeExplainer and the effective local approximations from LIME demonstrate that interpretability tools can scale to real-world applications while maintaining theoretical rigor. By prioritizing methodological demonstration over benchmark optimization, this research provides a foundational resource for researchers and practitioners seeking to implement explainable AI in their domains.

5.2 Recommendations

5.2.1 Practical Implementation

Organizations deploying machine learning models should integrate XAI techniques as standard components of their model development pipelines rather than optional additions. The

demonstrated workflow requires minimal computational resources and executes within minutes on standard hardware, making implementation feasible across diverse organizational contexts. Technical teams should prioritize establishing automated explanation generation for all production models, utilizing SHAP for global feature importance analysis and LIME for case-specific explanations to stakeholders.

Model validation processes should incorporate explanation consistency as an evaluation metric alongside traditional performance measures. The high correlation observed between SHAP and LIME suggests that significant disagreement between methods may indicate model instability or data quality issues requiring investigation. Documentation standards should mandate inclusion of both global feature importance rankings and representative local explanations for critical predictions, ensuring model behavior remains auditable and comprehensible.

Educational institutions teaching machine learning should adopt similar reproducible workflows to demonstrate explainability concepts. The complete pipeline from data preprocessing through explanation generation provides students with practical experience in implementing XAI techniques while reinforcing the importance of interpretability in responsible AI development. The accessibility of the implementation on CPU hardware ensures equitable access to these educational resources regardless of computational infrastructure.

5.2.2 Methodological Improvements

Future implementations should explore ensemble explanation approaches that combine multiple XAI methods to provide more robust interpretations. While this study demonstrated strong agreement between SHAP and LIME, systematic integration of their complementary strengths could yield more comprehensive explanations. Specifically, SHAP's global consistency could guide LIME's local sampling strategy, potentially improving approximation accuracy in complex feature spaces.

The explanation pipeline should be extended to handle diverse data modalities beyond tabular structures. Adaptation of the workflow for text, image, and time series data would broaden its applicability while maintaining the core principles of accessibility and reproducibility. Particular attention should focus on maintaining computational efficiency as data complexity increases, possibly through selective explanation generation for representative samples rather than exhaustive analysis.

Visualization techniques require enhancement to better communicate uncertainty in explanations. Current implementations present point estimates of feature importance without conveying confidence intervals or stability measures. Incorporating bootstrap-based confidence bands for SHAP values and displaying LIME's local fidelity scores would provide users with better calibrated trust in explanations.

5.2.3 Future Research Directions

Systematic investigation of explanation stability across different model architectures and hyperparameter configurations would strengthen understanding of XAI method reliability. Research should examine whether the high correlation observed between SHAP and LIME persists across gradient boosting machines, neural networks, and other model families, identifying conditions where explanation methods diverge and investigating underlying causes.

Development of standardized benchmarks for evaluating explanation quality remains a critical need. While this study used correlation between methods as a validation metric, comprehensive evaluation frameworks should incorporate human subject studies, counterfactual analysis, and downstream decision impact assessment. Such benchmarks would enable principled comparison of emerging XAI techniques and guide selection for specific application contexts.

Investigation of the relationship between model complexity, predictive performance, and explanation quality warrants deeper exploration. This study maintained interpretability through

architectural constraints, but systematic analysis of this trade-off across multiple datasets and domains would inform optimal model design strategies. Research should particularly focus on identifying minimal complexity thresholds required for capturing domain-specific patterns while preserving explainability.

The development of interactive explanation systems that allow stakeholders to explore model behavior through guided interrogation represents a promising direction. Static explanations, while valuable, cannot address all potential questions about model decisions. Dynamic interfaces enabling users to modify input features and observe explanation changes would enhance understanding and build appropriate trust in automated decisions.

Finally, research should address the scalability of XAI methods to large-scale production systems processing millions of predictions daily. While the demonstrated workflow handles moderate data volumes effectively, industrial applications require streaming explanation generation, efficient storage of explanation artifacts, and real-time anomaly detection based on explanation patterns. Solutions maintaining the accessibility demonstrated in this work while achieving enterprise scale would significantly advance practical XAI deployment.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 648-657). https://doi.org/10.1145/3351095.3375624

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. Bank of England Working Paper No. 816. https://doi.org/10.2139/ssrn.3435104

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2019). L-Shapley and C-Shapley: Efficient model interpretation for structured data. In International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1808.02610

Covert, I., Lundberg, S., & Lee, S. I. (2020). Understanding global feature contributions with additive importance measures. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (pp. 17212-17223). https://doi.org/10.48550/arXiv.2004.00668

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42. https://doi.org/10.1145/3236009

Kumar, E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In Proceedings of the 37th International Conference on Machine Learning (pp. 5491-5500). https://doi.org/10.48550/arXiv.2002.11097

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2021). Shapley residuals: Quantifying the limits of the Shapley value for explanations. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021) (pp. 26598-26608). https://doi.org/10.48550/arXiv.2106.14139

Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature Communications, 11(1), 3852. https://doi.org/10.1038/s41467-020-17431-x

Lipton, Z. C. (2018a). The mythos of model interpretability. Queue, 16(3), 31-57. https://doi.org/10.1145/3236386.3241340

Lipton, Z. C. (2018b). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Communications of the ACM, 61(10), 36-43. https://doi.org/10.1145/3233231

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 4765-4774). https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 607-617). https://doi.org/10.1145/3351095.3372850

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019a). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C. (2019b). Please stop explaining black box models for high stakes decisions. In *Proceedings of the Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1811.10154

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180-186). https://doi.org/10.1145/3375627.3375830

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319-3328). https://doi.org/10.48550/arXiv.1703.01365

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 10-19). https://doi.org/10.1145/3287560.3287566

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International