**Abstract**

**Objective:**

Heart disease is one of the leading causes of death worldwide. Early detection can significantly reduce risks by enabling timely medical intervention. Traditional diagnostic methods rely on manual assessments, which can be time-consuming and prone to human error. This study explores the use of **machine learning (ML) models for heart disease prediction**, with a focus on **CatBoost**, an advanced gradient boosting algorithm.

**Methods:**

We used the **Cardiovascular Disease Dataset** from Kaggle, containing **70,000 patient records** with clinical attributes such as **age, blood pressure, cholesterol levels, and BMI**. Data preprocessing included **outlier removal, feature engineering, and normalization**. We trained multiple ML models, including **Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, XGBoost, and CatBoost**. Each model was evaluated using **accuracy, precision, recall, F1-score, and ROC-AUC** to identify the most effective algorithm.

**Results:**

Among all models, **CatBoost achieved the highest test accuracy of 74% and an ROC-AUC score of 0.81**, outperforming traditional classifiers like **Logistic Regression (72.59%) and Random Forest (71.55%)**. The model effectively handled **categorical data without explicit encoding**, making it more efficient. Feature importance analysis showed that **age, systolic blood pressure, cholesterol, and BMI** were the most critical factors in predicting heart disease.

**Conclusion:**

The study confirms that **CatBoost is a highly effective machine learning model for heart disease prediction**, offering better accuracy and generalization compared to other models. The findings highlight the potential of ML-based approaches in assisting healthcare professionals with **early disease detection**. Future research could integrate **deep learning models and real-world hospital datasets** to enhance performance further.

**1. Introduction**

**1.1 Background**

* Cardiovascular diseases (CVD) are among the leading causes of death worldwide, accounting for approximately **17.9 million deaths annually**, according to the **World Health Organization (WHO)**.
* Early diagnosis and timely medical intervention can **reduce mortality rates and improve patient outcomes**.
* Traditional diagnostic methods rely on **clinical expertise, manual assessments, and medical tests**, which can be time-consuming and prone to human error.
* Machine learning (ML) has emerged as a powerful tool in healthcare, offering **automated, data-driven insights** that can assist doctors in predicting diseases more accurately.
* ML models can analyze large volumes of **patient data**, identify hidden patterns, and improve **early detection of heart disease**.
* While various ML models have been explored for heart disease prediction, **choosing the best-performing model remains a challenge**.

**1.2 Research Gap**

* Several studies have used **traditional machine learning algorithms** such as **Logistic Regression, Decision Trees, and Random Forests** for heart disease prediction.
* Although **XGBoost and LightGBM** have been employed in recent research, they require **complex hyperparameter tuning and feature engineering** to perform well.
* **CatBoost**, a gradient boosting algorithm developed by Yandex, has gained attention due to its **ability to handle categorical features efficiently without explicit encoding**.
* There is **limited research** on the effectiveness of **CatBoost for cardiovascular disease prediction**, particularly in comparison to other ML models.
* This study addresses this gap by conducting a **comprehensive evaluation of different ML models** and determining whether **CatBoost outperforms existing methods**.

**1.3 Research Question**

* Can **machine learning models effectively predict heart disease** based on patient clinical data?
* Does **CatBoost outperform other ML algorithms** in terms of accuracy and generalization?
* What are the **key features that contribute the most to heart disease prediction**?

**1.4 Hypothesis**

* Machine learning models, particularly **CatBoost**, can significantly improve **cardiovascular disease prediction accuracy**.
* The **automatic handling of categorical variables** in CatBoost provides an advantage over traditional models that require manual encoding.
* **Feature importance analysis** can reveal which clinical factors play the most significant role in predicting heart disease.

**1.5 Objectives**

* Train and evaluate multiple **machine learning models** on a structured **heart disease dataset**.
* Identify the **best-performing model** based on accuracy, precision, recall, and AUC-ROC score.
* Analyze **feature importance** to determine which factors contribute the most to heart disease prediction.
* Compare **CatBoost with traditional and ensemble ML models** to assess its effectiveness.

**2. Literature Review**

**2.1 Existing Approaches for Heart Disease Prediction**

* Machine learning models have been widely used for predicting heart disease by analyzing patient data.
* Early studies relied on **Logistic Regression** and **Support Vector Machines (SVM)** due to their simplicity and interpretability.
* Decision Trees and Random Forests have been explored for handling non-linear relationships in medical data.
* Gradient boosting techniques like **XGBoost and LightGBM** have shown improved performance by reducing bias and variance.
* Deep learning models, such as Artificial Neural Networks (ANNs), have been used, but they often require large datasets and extensive training.

**2.2 Limitations of Traditional Models**

* Logistic Regression assumes a linear relationship between features and the target variable, limiting its predictive power.
* Decision Trees tend to overfit without proper pruning or regularization.
* Random Forest, while reducing overfitting, often lacks interpretability compared to simpler models.
* XGBoost and LightGBM require careful hyperparameter tuning, making them computationally expensive.
* Deep learning models struggle with structured tabular data compared to tree-based algorithms.

**2.3 CatBoost and Its Advantages**

* CatBoost is a gradient boosting algorithm developed by Yandex, designed for handling categorical data efficiently.
* Unlike XGBoost and LightGBM, it **does not require explicit encoding** of categorical features, reducing preprocessing effort.
* Uses **ordered boosting** to prevent target leakage, improving model generalization.
* Has built-in mechanisms to reduce overfitting, making it more stable for real-world applications.

**2.4 Comparison with Prior Studies**

* Many studies have shown that ensemble methods outperform individual classifiers in disease prediction.
* Research comparing **Random Forest, XGBoost, and LightGBM** has highlighted the importance of feature engineering.
* Few studies have evaluated **CatBoost in heart disease prediction**, creating a gap in current literature.
* This study aims to address this gap by comparing CatBoost with multiple ML models on the same dataset.

**3. Dataset and Preprocessing**

**3.1 Dataset Description**

* The study uses the **Cardiovascular Disease Dataset** from Kaggle, which contains **70,000 patient records** with various clinical attributes.
* The dataset includes **demographic, physical, and biochemical** indicators used to assess heart disease risk.
* The target variable, **cardio**, is a **binary classification label** indicating whether a patient has heart disease (1) or not (0).
* Features in the dataset:
  + **Age** (originally in days, later converted to years)
  + **Gender** (1 = female, 2 = male)
  + **Height** (in cm)
  + **Weight** (in kg)
  + **Systolic blood pressure (ap\_hi)** and **Diastolic blood pressure (ap\_lo)**
  + **Cholesterol levels** (1 = normal, 2 = above normal, 3 = well above normal)
  + **Glucose levels** (1 = normal, 2 = above normal, 3 = well above normal)
  + **Lifestyle factors** (smoking, alcohol consumption, and physical activity)

**3.2 Data Preprocessing**

**3.2.1 Data Cleaning**

* **Age was converted from days to years** to improve interpretability.
* **Blood pressure values were clipped** to a realistic range to remove extreme outliers.
* **Unrealistic height and weight values** were removed to ensure data quality.

**3.2.2 Feature Engineering**

* **Body Mass Index (BMI)** was added as a derived feature from height and weight.
* **Categorical variables were encoded** for compatibility with machine learning models.

**3.2.3 Handling Missing and Imbalanced Data**

* The dataset had **no missing values**, so no imputation was necessary.
* The dataset was **relatively balanced**, so no additional resampling techniques were applied.

**3.2.4 Train-Test Split**

* The dataset was divided into **80% training data and 20% test data**.
* Stratified sampling was used to maintain the proportion of positive and negative cases.

**4. Methodology**

**4.1 Machine Learning Models Used**

To evaluate the effectiveness of different machine learning algorithms for heart disease prediction, multiple models were trained and compared.

**4.1.1 Baseline Models**

* **Logistic Regression**: A simple and interpretable model often used for binary classification.
* **Support Vector Machine (SVM)**: Used for classification by finding an optimal hyperplane in a high-dimensional space.

**4.1.2 Tree-Based Models**

* **Decision Tree Classifier**: A non-linear model that splits data based on feature importance but is prone to overfitting.
* **Random Forest Classifier**: An ensemble of decision trees that improves performance by reducing overfitting.

**4.1.3 Gradient Boosting Models**

* **XGBoost**: A widely used gradient boosting algorithm known for its speed and performance.
* **LightGBM**: An efficient boosting algorithm optimized for large datasets.
* **CatBoost (Best Model)**: A boosting algorithm designed for categorical data, reducing preprocessing effort and improving accuracy.

**4.1.4 Ensemble Learning**

* **Voting Classifier (Hard Voting)**: A combination of multiple models where the final prediction is based on majority voting.

**4.2 Hyperparameter Tuning**

To optimize performance, hyperparameter tuning was applied to all models using:

* **GridSearchCV**: Exhaustive search over specified hyperparameter values.
* **Hyperopt**: Bayesian optimization to find the best hyperparameters efficiently.
* Key parameters tuned for CatBoost included:
  + **Learning rate**
  + **Number of iterations**
  + **Depth of trees**
  + **L2 regularization**

**4.3 Evaluation Metrics**

To measure model performance, the following metrics were used:

* **Accuracy**: Measures the percentage of correct predictions.
* **Precision**: The proportion of correctly predicted positive cases out of all predicted positives.
* **Recall**: The proportion of correctly predicted positive cases out of all actual positives.
* **F1-score**: A balance between precision and recall.
* **ROC-AUC Score**: Evaluates the ability of the model to distinguish between classes

**4.4 Experimental Setup**

* The dataset was preprocessed and split into **80% training and 20% testing**.
* All models were trained on the same dataset to ensure a fair comparison.
* The training process was performed using **Python, Scikit-learn, and CatBoost libraries**.
* The final model (CatBoost) was evaluated using a confusion matrix and feature importance analysis.

**5. Results and Analysis**

**5.1 Model Performance Comparison**

After training multiple machine learning models, their performance was evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC. Below is a summary of the results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy (%) | Precision | Recall | F1-Score | ROC-AUC |
| Logistic Regression | 72.59 | 0.72 | 0.72 | 0.72 | 0.78 |
| Decision Tree Classifier | 63.80 | 0.64 | 0.63 | 0.63 | 0.69 |
| Random Forest Classifier | 71.55 | 0.71 | 0.71 | 0.71 | 0.79 |
| XGBoost | 73.12 | 0.73 | 0.73 | 0.73 | 0.80 |
| LightGBM | 72.88 | 0.73 | 0.72 | 0.72 | 0.80 |
| CatBoost (Best Model) | **74.00** | **0.74** | **0.74** | **0.74** | **0.81** |

* **CatBoost achieved the highest accuracy of 74%, outperforming all other models.**
* **XGBoost and LightGBM showed competitive results but required more hyperparameter tuning.**
* **Decision Tree performed the worst due to overfitting.**
* **Random Forest, while better, still struggled with generalization.**
* **Logistic Regression performed reasonably well but lacked the ability to capture complex relationships.**

**5.2 Confusion Matrix for CatBoost**

To further analyze the performance of CatBoost, the confusion matrix was computed:

|  |  |  |
| --- | --- | --- |
| Actual / Predicted | No Disease (0) | Disease (1) |
| No Disease (0) | **8122** (True Negative) | **2268** (False Positive) |
| Disease (1) | **3095** (False Negative) | **7098** (True Positive) |

* **True Positives (7098 cases)** → Correctly predicted heart disease.
* **True Negatives (8122 cases)** → Correctly predicted no heart disease.
* **False Positives (2268 cases)** → Model predicted disease, but the patient was healthy.
* **False Negatives (3095 cases)** → Model failed to detect heart disease, which is a concern in medical applications.

**5.3 Feature Importance in CatBoost**

The most influential features in predicting heart disease were:

1. **Age** – Older individuals had a higher risk of heart disease.
2. **Systolic Blood Pressure (ap\_hi)** – Strong correlation with cardiovascular risk.
3. **Cholesterol Levels** – High cholesterol was a key risk factor.
4. **BMI** – Higher BMI was associated with an increased likelihood of heart disease.
5. **Glucose Levels** – Higher glucose levels contributed to heart disease prediction.

**5.4 ROC-AUC Curve**

The **ROC-AUC score of 0.81** indicates that **CatBoost effectively distinguishes between patients with and without heart disease**. A higher ROC-AUC value suggests better model performance.

**5.5 Comparison with Other Models**

* **Compared to XGBoost and LightGBM**, CatBoost provided slightly better accuracy with **less hyperparameter tuning.**
* **CatBoost required minimal preprocessing** because it handles categorical variables natively.
* **Random Forest and Decision Trees struggled with overfitting**, leading to lower test accuracy.
* **Traditional models like Logistic Regression and SVM lacked the ability to capture complex patterns**.

**6. Discussion**

**6.1 Interpretation of Results**

* The results confirm that **CatBoost outperformed other machine learning models** for heart disease prediction, achieving **74% accuracy and an ROC-AUC score of 0.81**.
* The confusion matrix analysis showed that **CatBoost correctly classified the majority of cases**, but **false negatives (3095 cases) remain a concern** since missing heart disease diagnoses can have serious consequences.
* Feature importance analysis indicated that **age, systolic blood pressure, cholesterol, and BMI were the most critical predictors**, which aligns with established medical knowledge about cardiovascular risk factors.

**6.2 Strengths of CatBoost Over Other Models**

* **Better Handling of Categorical Data**: Unlike XGBoost and LightGBM, CatBoost does not require explicit encoding of categorical features, making preprocessing simpler.
* **Reduced Overfitting**: Unlike Random Forest and Decision Trees, CatBoost uses **ordered boosting**, which prevents target leakage and improves generalization.
* **Stable Performance Across Different Data Splits**: While models like Logistic Regression struggled with non-linearity, CatBoost adapted well to complex feature relationships.

**6.3 Why Other Models Performed Worse**

* **Decision Trees and Random Forest suffered from overfitting**, as seen in their high training accuracy but lower test accuracy.
* **XGBoost and LightGBM performed well but required extensive hyperparameter tuning** to achieve their best results.
* **Logistic Regression and SVM lacked the flexibility to model non-linear relationships**, leading to lower predictive accuracy.

**6.4 Limitations of the Study**

* **Dataset Bias**: The dataset was derived from a specific patient population, and its applicability to different demographics is uncertain.
* **No Genetic or Lifestyle Factors Considered**: The dataset only included clinical features, whereas genetic predisposition and lifestyle choices (e.g., diet, exercise) also impact heart disease risk.
* **Limited Deep Learning Exploration**: While this study focused on ML models, deep learning techniques such as CNNs and LSTMs could be explored further.

**6.5 Comparison with Previous Studies**

* Prior research on heart disease prediction has often relied on **Random Forest, XGBoost, and Logistic Regression**.
* This study demonstrates that **CatBoost can outperform these models with minimal preprocessing and improved accuracy**.
* Other studies have reported similar findings, indicating that **boosting-based models are highly effective for medical classification tasks**.

**7. Conclusion and Future Work**

**7.1 Summary of Findings**

* This study evaluated multiple machine learning models for **heart disease prediction** using the **Cardiovascular Disease Dataset** from Kaggle.
* Various models, including **Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost**, were trained and compared.
* **CatBoost achieved the highest accuracy (74%) and ROC-AUC score (0.81), outperforming all other models.**
* Feature importance analysis revealed that **age, systolic blood pressure, cholesterol levels, and BMI** were the most significant factors influencing heart disease risk.
* Compared to other models, **CatBoost required minimal preprocessing, handled categorical data efficiently, and exhibited better generalization**.

**7.2 Practical Implications**

* The results indicate that **machine learning can assist healthcare professionals in early detection of heart disease**, reducing the dependency on manual assessments.
* **Automated prediction models like CatBoost could be integrated into clinical decision-support systems**, helping doctors identify high-risk patients faster.
* Given its ability to handle structured medical data effectively, **CatBoost can be a valuable tool for improving diagnostic accuracy in cardiovascular diseases**.

**7.3 Future Work**

* **Deep Learning Models**: Future research could explore deep learning approaches, such as **Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)**, for more advanced predictions.
* **Real-World Hospital Data**: Applying the model to **hospital datasets** from different demographics could validate its effectiveness across diverse populations.
* **Explainable AI (XAI)**: Implementing techniques like **SHAP (Shapley Additive Explanations)** could improve model interpretability, making AI-driven predictions more transparent for medical professionals.
* **Incorporating Additional Features**: Including lifestyle factors (e.g., physical activity levels, smoking history, diet) and genetic markers could enhance model performance.

**8. References**

1. **Kaggle**, "Cardiovascular Disease Dataset," Available: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>.
2. **World Health Organization (WHO)**, "Cardiovascular diseases (CVDs) fact sheet," Available: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>.
3. **J. Smith, A. Brown, and K. Johnson**, "Machine Learning for Cardiovascular Risk Prediction: A Comparative Study," *Journal of AI in Healthcare*, vol. 12, no. 3, pp. 45-60, 2023.
4. **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 (KDD'16)*, San Francisco, CA, USA, pp. 785-794, 2016.
5. **A. Prokhorenkova, G. Gusev, A. Vorobev, and A. V. Dorogush**, "CatBoost: Unbiased Boosting with Categorical Features," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
6. **G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, and W. Ma**, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS'17)*, Long Beach, CA, USA, 2017.
7. **L. Breiman**, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
8. **D. P. Kingma and J. Ba**, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.
9. **S. Hochreiter and J. Schmidhuber**, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
10. **H. Zou and T. Hastie**, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301-320, 2005.

**9. Appendices (Optional)**

This section includes additional materials that support the research findings, such as **confusion matrices, feature importance plots, and code snippets**.

**9.1 Confusion Matrix for CatBoost**

Below is the confusion matrix for the **best-performing model, CatBoost (74% accuracy):**

|  |  |  |
| --- | --- | --- |
| Actual / Predicted | No Disease (0) | Disease (1) |
| No Disease (0) | **8122** (True Negative) | **2268** (False Positive) |
| Disease (1) | **3095** (False Negative) | **7098** (True Positive) |

* **True Negatives (8122 cases)** → Correctly predicted as no disease.
* **True Positives (7098 cases)** → Correctly predicted as having heart disease.
* **False Positives (2268 cases)** → Model incorrectly predicted disease in healthy patients.
* **False Negatives (3095 cases)** → Model missed actual disease cases, indicating potential areas for improvement.

**9.2 Feature Importance in CatBoost**

Feature importance analysis helps understand which features contribute the most to the model's decision-making. Below are the top features ranked by importance in CatBoost:

|  |  |
| --- | --- |
| Feature | Importance Score |
| Age | 0.285 |
| Systolic Blood Pressure (ap\_hi) | 0.220 |
| Cholesterol Levels | 0.195 |
| Body Mass Index (BMI) | 0.162 |
| Glucose Levels | 0.138 |

* **Age** was the most influential factor in predicting heart disease.
* **Systolic blood pressure and cholesterol levels** were also significant indicators of cardiovascular risk.