JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Big Data-Driven Health Risk Stratification: A Health Index-Based Approach Using Feature Importance and PySpark

Oluwasegun Abiodun Abioye [1,*] and Martins Ekata Irhebhude [2]

[1] Directorate Of Information and Communications Technology, Nigerian Defence Academy, Kaduna Nigeria; e-mail : segunabioye@nda.edu.ng
[2] Department of Computer Science, Nigerian Defence Academy, Kaduna Nigeria; e-mail : mirhebhude@nda.edu.ng
* Corresponding Author : Oluwasegun Abiodun Abioye

Abstract: Health risk stratification is crucial for preventive healthcare, yet existing models often rely on binary classification generalized disease prediction, neglecting personalized health indicators and graded risk levels. Many studies apply feature selection techniques like Relief and Univariate Selection without quantifying the weighted impact of features. To address these gaps, this study introduces a Big Data-driven Health Index (HI) framework using PySpark for scalable health risk stratification. The HI is computed as a weighted sum of health-related features using SHAP Analysis, XGBoost, Random Forest, and Correlation Analysis. PySpark enables efficient processing of large-scale health data, and individuals are classified into Low and High Risk. Optimal classification thresholds are determined using the Youden Index from the ROC curve to balance sensitivity and specificity. Personalized health recommendations are generated based on risk categories to guide preventive interventions. Performance evaluation reveals that Correlation Analysis achieves 100% precision and 98.90% recall, outperforming other methods. SHAP prioritizes recall but has low precision, while XGBoost and Random Forest improve precision but struggle with recall. By leveraging Big Data techniques with PySpark, this study enhances computational efficiency, scalability, and classification accuracy, addressing prior research limitations and providing a robust data-driven approach to personalized health monitoring.

Keywords: Big Data Analytics; Feature Importance; Health Index; Heart Disease Risk; PySpark; Risk Stratification.

1. Introduction

Information that cannot be processed or stored using conventional methods is called Big Data[1]. However, traditional systems can no longer accommodate the immense scale of data being generated, necessitating a more advanced framework comprising multiple components that perform specialized tasks. Big Data has five key attributes: volume, variety, velocity, value, and veracity [2]. The rapid technological advancement in applicative areas such as the IoT, cloud computing, and edge computing combined with extensive use by society has resulted in massive overflows of data generated every fraction of a second globally [3].

In the health sector, this explosion of data is particularly significant, as it includes patient records, diagnostic imaging, genetic information, wearable health device data, and real-time monitoring from IoT devices. Big data in healthcare and medicine consists of large and complex datasets[4], including the heart disease dataset analyzed in this research. Medical data's vast volume and complexity hinder effective analysis and restrict its practical application in clinical settings[5]. Health data sources have evolved to include computerized physician order entries, electronic medical records (EMRs), clinical notes, medical images, cyber-physical systems, the medical Internet of Things (IoT), genomic data, and clinical decision support systems. Traditional EMR-based software and hospital informatics systems are insufficient for efficiently managing and analyzing healthcare datasets[6]. However, this data's size, unstructured nature, and abstract form present challenges for healthcare organizations in collecting, analyzing, and interpreting it effectively[7]. Overcoming these challenges can enable

Received: February, 5th 2025
Revised: March, 17th 2025
Accepted: March, 19th 2025
Published: March, 24th 2025

healthcare professionals to make timely, data-driven decisions, predict disease outbreaks, enhance treatment personalization, and improve overall health outcomes.

Evaluating big data requires the use of appropriate analytical tools[8]. The growing need to manage the ever-increasing volume of data has driven significant interest in developing effective big data frameworks [9]. Extensive research has explored various aspects of big data, including infrastructure[10], management [11], data searching [12], mining [13], and security [14]. Big data infrastructures have been designed to support analytics with fast, reliable, and adaptable computational architectures, offering efficient quality attributes such as flexibility, accessibility, resource pooling, and ease of use on-demand[15]. The importance of an effective big data analysis framework becomes evident when processing algorithms on extensive datasets [16]. While local systems typically rely on a single central processing unit (CPU), the increasing size of datasets has led to the growing adoption of multi-core graphics processing units (GPUs) to boost performance. Although parallel processing can be efficiently implemented due to distributed architectures, GPUs are often cost-prohibitive or unavailable. Therefore, there is a persistent demand for tools that leverage accessible CPUs in a distributed manner within local systems[17].

PySpark is Python's interface to the programming interface of Apache Spark, an open-source distributed computing system intended for processing and analyzing big data [18]. It allows developers to scale and speed their application with Python's simplicity and library ecosystem [19]. Some of its major components include SparkContext, which is the entry into Spark functionality by coordinating an operation across the cluster; Resilient Distributed Dataset (RDD), a fault-tolerant, distributed data structure; and DataFrames and SQL, which provides a much-used Application Programming Interface (API) for structured records process and query [20]. One major benefit PySpark offers is its capacity to run processes in memory, which promises high performance while processing large-scale datasets [21]. Despite some advantages, some disadvantages include learning to configure clusters and debugging distributed applications. While many such strengths and challenges for PySpark make it appealing, it is quite possible to say that PySpark fits in well with most modern big data projects, ranging from different choices[22].

This paper proposes personal health indicators to compute health index (HI), categorize individuals into risk levels, and generate personalized health recommendations using a big data health-related dataset of 319,797 instances and 18 features. A set of calculated weights is assigned to features, indicating their contribution to health or associated risks. Using these weights, a HI is computed for each individual by summing the weighted values of the features. Based on the HI, individuals are categorized into Low and High Risk levels. The categorization uses conditional logic, with thresholds set for the index values.

Using a HI with static thresholds instead of a Machine Learning (ML)-based classifier offers several advantages, particularly in system monitoring, predictive maintenance, and health assessment. It ensures simplicity and transparency, as fixed thresholds provide clear decision-making criteria without the complexity of ML models, which often function as black boxes[23]. This transparency enhances trust and ease of interpretation, making it valuable in fields requiring explainability [24]. Additionally, consistency and predictability are key benefits, as HI-based methods maintain stable decision boundaries, unlike ML classifiers that may shift based on training data variations [25]. Another advantage is independence from training data, reducing the risk of performance degradation due to incomplete or biased datasets[26]. They also reduce the risk of overfitting, as they rely on predefined rules rather than data-driven generalizations, ensuring robustness across diverse scenarios[27]. However, a key trade-off is that static thresholds assume universal, predefined limits, which may not capture nuanced or evolving patterns as effectively as adaptive ML models[25]. Ultimately, HI-based methods are most effective when explainability, reliability, and efficiency precede adaptability, particularly in critical systems like industrial equipment and medical devices [28].

The structure of this research paper is as follows: Section 2 reviews previous studies and key concepts in big data analytics, particularly in the context of PySpark, feature importance analysis, and machine learning applications for health risk assessment. Section 3 describes the proposed methodology, detailing data preprocessing, feature selection, health index computation, and classification techniques. Section 4 presents the experimental setup and results, including dataset description, correlation analysis, classification performance comparison, and personalized health recommendations. Section 5 discusses the key findings, emphasizing the

significance of the proposed framework, analyzing its implications, and outlining directions for future research.

2. Related Work

Several studies have explored machine learning-based approaches for heart disease prediction and risk assessment, often leveraging big data analytics. Study [29] developed a real-time heart disease prediction system that integrates Apache Spark for streaming big data processing and MLlib for classification, achieving 87.50% accuracy. The model demonstrated effective real-time monitoring using Spark Streaming and Apache Cassandra for scalable data handling. Similarly, [30] proposed an ECG classification system using Spark-Scala and MLlib, achieving high accuracy (96.75% with GDB and 97.98% with Random Forest). However, both studies focus on binary classification without incorporating personalized health indicators.

Research [31] introduced a real-time arrhythmia detection pipeline using Apache Spark's Structured Streaming module, demonstrating an accuracy of 88.7% with a Random Forest classifier. The study highlights the efficiency of Spark-based frameworks for real-time cardiac monitoring but does not quantify the impact of specific health features on classification outcomes. Study [32] developed Sehaa, a healthcare analytics tool using Apache Spark and Twitter data, employing Naïve Bayes and Logistic Regression for classifying health-related tweets. Although it provides insights into public health trends, it lacks an individualized risk assessment approach.

Several studies have also explored feature selection techniques for heart disease prediction. For example, research [33] utilized Univariate Feature Selection and Relief to select relevant features before training models such as Decision Tree, SVM, and Random Forest, achieving a maximum accuracy of 94.9%. Another study[34] proposed a real-time disease prediction system using streaming data from Twitter and Kafka, evaluating various ML models, with Random Forest achieving the highest classification accuracy (92.05%). However, these studies focus on disease classification without implementing a comprehensive health risk stratification framework.

Further, deep learning methods have been employed in cardiac diagnosis. Research [6] applied transfer learning on Apache Spark for ECG image classification, utilizing InceptionV3 and Logistic Regression. Although this approach enhances classification performance, it does not provide interpretability or feature weighting insights. Studies [35] and [36] explored big data-driven cardiovascular disease prediction, integrating feature fusion and hybrid deep learning models. However, these studies lack a unified health assessment and risk stratification index.

Integrating HI and PySpark presents a notable advancement over prior methods in healthcare analytics. While previous studies primarily focus on disease classification or real-time monitoring, they often lack personalized health assessment and interpretable insights. The HI addresses this gap by providing individualized risk evaluations through a weighted summation of selected features. Concurrently, PySpark facilitates scalable and real-time data processing, enabling healthcare systems to monitor and assess thousands of individuals simultaneously [37]. Its dynamic data handling capabilities allow for continuous updates to the HI, supporting timely risk stratification and targeted interventions[38]. Moreover, the combination of interpretable feature weighting and big data processing ensures that healthcare providers obtain clear, actionable insights, ultimately enhancing clinical decision-making and improving patient outcomes[39].

Despite advancements in real-time disease prediction and feature selection methods, existing studies exhibit gaps in personalized health risk assessment, graded risk stratification, and feature impact quantification. The proposed study introduces a HI-based framework that categorizes individuals into low-risk and high-risk groups, leveraging feature weighting techniques such as Correlation Analysis, SHAP, XGBoost, and Random Forest. Unlike previous methods, this approach ensures interpretability and scalability using PySpark, facilitating efficient big data processing for health risk stratification (see Table 1).

Table 1 compares previous methods with the proposed HI + PySpark framework, highlighting its advantages in personalization, risk stratification, scalability, interpretability, and integration. While previous studies rely on black-box ML models and separate data processing

tools, the proposed approach offers an end-to-end workflow using PySpark, enabling a transparent, scalable, and data-driven health monitoring system.

Table 1. Depicts the Comparison with the Previous Method and the Proposed Method.

Aspects	Previous Methods	Health Index + PySpark (Ours)
Personalization	Limited to binary or disease-specific predictions.	Holistic assessment using multiple health indicators.
Risk Stratification	Often absent or limited.	Classifies individuals into Low, Moderate, High-risk levels.
Scalability	Struggles with large-scale data.	Efficiently processes big data using distributed computing.
Interpretability	Black-box models with limited transparency.	Transparent weighted feature summation.
Integration	Separate tools for data processing and Machine Learning.	End-to-end workflows with PySpark and MLlib.

3. Proposed Method

This section provides an overview of the experimental setup employed to develop the proposed models effectively. Certain important terminologies are explained in this section for optimal understanding. Figure 1 shows an overview of the proposed method.

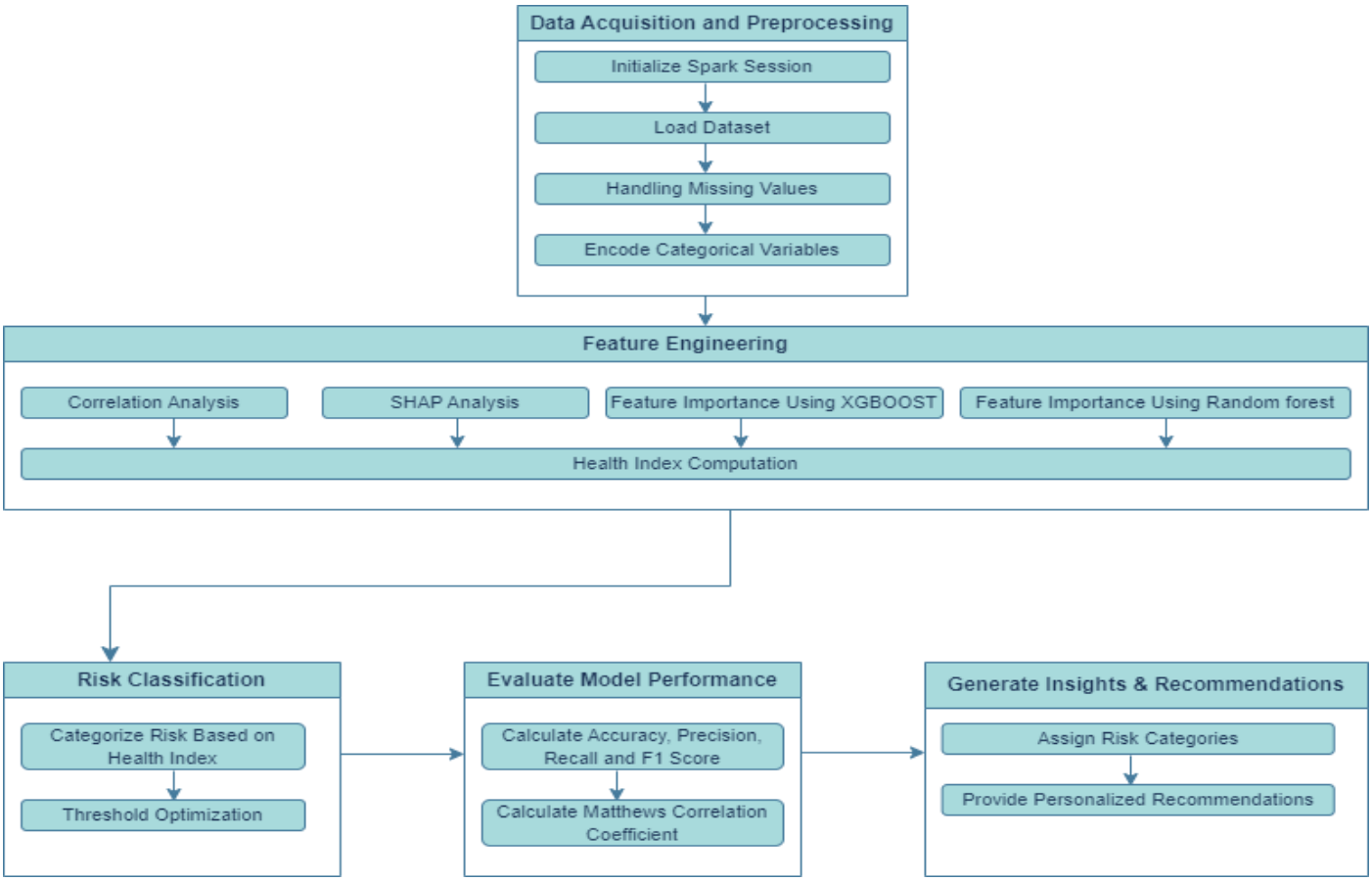

Figure 1. Methodology for Assessing Health Risk and Generating Personalized Health Recommendations

3.1. PySpark Data Processing

The PySpark framework is the Python API for Apache Spark, an open-source distributed computing system designed for large-scale data processing. It enables Python developers to harness the full power of Apache Spark for big data processing, real-time analytics, and machine learning while benefiting from Spark's speed and scalability. PySpark supports in-

memory computing, which significantly improves performance by processing data in RAM rather than relying on traditional disk-based systems [21]. The framework provides high-level APIs for batch and stream processing, Resilient Distributed Datasets (RDDs), and data frames for efficient distributed data handling. PySpark also integrates with MLlib, Spark's machine learning library, offering various classification, regression, and clustering algorithms while supporting real-time data processing via Structured Streaming [40]. PySpark's ability to integrate with other big data technologies such as Hadoop, Hive, and Kafka makes it suitable for various applications, including data transformation, ETL processes, and large-scale machine learning. Its flexibility allows deployment across various cluster managers, including YARN, Mesos, and Kubernetes, providing ease of use on multiple platforms [41]. This makes PySpark ideal for distributed data processing, machine learning, and real-time analytics[21]. The following sections describe how PySpark's DataFrames, parallel processing, caching, and ML libraries are applied in the study.

3.1.1. Use of DataFrames

PySpark's DataFrame API enables efficient, structured data handling by supporting SQL-like operations such as filtering, aggregation, and transformation[42]. This study loads the dataset as a data frame, ensuring optimized query execution and scalability. This structured approach allows seamless integration with PySpark's MLlib, facilitating feature engineering, model training, and evaluation.

3.1.2. Parallel Processing with RDDs

PySpark distributes computations across multiple nodes using Resilient Distributed Datasets (RDDs), which enable parallel execution of transformations such as filtering and feature selection [43]. This distributed computing framework significantly improves the performance of large-scale healthcare data processing.

3.1.3. Feature Engineering & Caching

To enhance computational efficiency, Vector Assembler is used to construct feature vectors, consolidating multiple attributes into a single feature representation. Additionally, intermediate DataFrames are cached to avoid redundant computations, improving processing speed and scalability [44].

3.1.4. Handling Missing Values Efficiently

Missing data is handled using mean imputation for numerical attributes and model-based imputation for categorical attributes. PySpark processes these operations in parallel, allowing efficient large-scale data handling [45]. This ensures data integrity while reducing computational overhead.

3.1.5. Machine Learning with PySpark

PySpark's MLlib library supports scalable machine learning algorithms. In this study, the RandomForestClassifier is utilized to analyze the importance of features and perform classification tasks efficiently [44]. This approach leverages PySpark's distributed computing capabilities to manage large datasets effectively.

3.1.6. Correlation Analysis for Feature Selection

Feature correlation is computed using PySpark's Correlation module, enabling efficient feature selection [46]. This step identifies the most significant predictors for health risk stratification, improving model interpretability and overall predictive performance. This study also compared several other feature importance analysis calculations, such as SHAP, XGBoost, and Random Forest.

3.2. Health Index Calculation

The HI is proposed as a method for quantifying an individual's health risk based on a weighted sum of selected health-related features. Each feature is assigned a weight that reflects its relative importance in predicting heart disease risk. The calculation follows the Equation (1).

$$HealthIndex = \sum_{i=1}^{n} (Feature_i \times Weight_i) \quad (1)$$

Where n represents the total number of features included in the model; $Feature_i$ is the observed value of the i-th health-related attribute for an individual; $Weight_i$ denotes the assigned weight for each feature, representing its contribution to heart disease risk assessment.

The feature weights are derived through a correlation-based weighting method, where each numerical feature's correlation with the target variable (heart disease) is calculated. Positive and negative correlations are considered absolute, ensuring that the most influential features are highly important regardless of direction. The absolute correlation values are summed to normalize the feature weights, and each feature's individual correlation is divided by this total. This ensures that all feature weights sum to one, allowing for a proportional representation of each feature's predictive power. Features with stronger correlations receive higher weights, indicating their more significant role in assessing health risk.

This method enables a data-driven and interpretable approach to health risk stratification, where features contributing positively to health are assigned positive weights, while risk factors receive negative weighting. The calculated HI is then used for risk classification, categorizing individuals into different health risk levels. The classification threshold is determined through statistical techniques, ensuring optimal separation between low-risk and high-risk groups.

3.3. Point-Biserial Correlation

The point-biserial correlation (r_{pb}) is equivalent to Pearson's product-moment correlation in cases where one variable is dichotomous (binary), represented by values 0 and 1, and the other variable is metric (measured on an interval or ratio scale)[47]. The point-biserial correlation coefficient (r_{pb}) calculated using Equation (2).

$$r_{pb} = \frac{M_1 - M_0}{s} . \sqrt{p \, . (1 - p)} \tag{2}$$

Where M_1 is mean of the continuous variable for the group coded as 1 in the binary variable; M_0: Mean of the continuous variable for the group coded as 0 in the binary variable; s: Standard deviation of the continuous variable; p: Proportion of the binary variable coded as 1.

The value of the point-biserial correlation coefficient (r_{pb}) ranges from -1 to 1, where r_{pb} = 1 indicates a perfect positive relationship, r_{pb}= −1 indicates a perfect negative relationship, and r_{pb}= 0 indicates no relationship. The sign of r_{pb} indicates the direction of the relationship: a positive value means that higher values of the continuous variable are associated with the group coded as 1 in the binary variable, while a negative value means that higher values of the continuous variable are associated with the group coded as 0[48]. This study uses the Point-biserial correlation coefficient to measure the strength and direction of the relationship between HI and the features with binary outcomes.

3.4. Youden's Index

Receiver operating characteristics (ROC) curves are utilized in biomedical research to assess the ability of biomarkers to differentiate between individuals with and without a disease[49]. The Youden index (J), defined as a function of sensitivity ($sensitivty(c)$) and specificity ($specficity(c)$), is widely used to assess overall diagnostic performance[50]. The index varies from 0 to 1, where values near 1 suggest the biomarker has high effectiveness, and values near 0 indicate minimal effectiveness. J is determined by Equation (3).

$$J = maximum \{sensitivty(c) + specficity(c) - 1\} \tag{3}$$

Where c is the cutoff for sensitivity and specificity. The Youden Index is used in this study to determine the optimal classification threshold for the HI, ensuring that the trade-off between sensitivity and specificity is optimized for accurate disease risk assessment.

3.5. Evaluations

The proposed HI framework undergoes a rigorous evaluation process to ensure its reliability in health risk stratification. The methodology incorporates multiple performance metrics to assess the classification accuracy and robustness of the model. The evaluation follows

a structured approach, beginning with the Receiver Operating Characteristic (ROC) curve analysis, which is employed to determine the model's ability to differentiate between low-risk and high-risk individuals. The optimal classification threshold is identified using Youden's Index, ensuring a balanced trade-off between sensitivity and specificity.

A confusion matrix is generated to validate the classification performance further, providing insight into the distribution of correctly and incorrectly classified instances across risk categories. The model's performance is quantified using key metrics such as accuracy, precision, recall, F1-score, and the Matthews Correlation Coefficient (MCC). These metrics offer a comprehensive evaluation, capturing both the correctness of predictions and the balance between false positives and false negatives.

The SHI distribution is also analyzed to examine the overall spread of risk scores within the dataset. The distribution pattern is expected to reveal underlying clusters, indicating varying degrees of health risk across individuals. By visualizing SHI scores, the methodology ensures that risk classification thresholds are optimally set, enabling effective differentiation between individuals requiring medical intervention and those maintaining stable health conditions.

3.6. Generating Insights and Recommendations

Following risk classification, the HI framework is designed to generate meaningful insights and personalized health recommendations. The insights derived from the model enable individuals to be assigned to distinct risk categories, guiding the development of personalized health interventions. The framework is structured to provide targeted recommendations based on the individual's classification, ensuring that high-risk individuals receive actionable guidance tailored to their specific health conditions. The recommendation system is designed to be adaptable, allowing for integration into healthcare decision-making processes where clinicians and healthcare professionals can leverage the HI to prioritize at-risk populations and formulate preventive health strategies.

4. Results and Discussion

4.1. Dataset

The Key Indicators of Heart Disease dataset is based on the CDC's 2020 annual survey, which collected health-related data from over 300,000 adults. Heart disease remains a leading cause of death in the U.S. across various racial groups, including African Americans, American Indians, Alaska Natives, and whites. The dataset contains 319,795 instances and 18 features, providing extensive health-related information for analyzing factors contributing to heart disease. The target variable, HeartDisease, is binary (Yes or No), indicating the presence of heart disease. Predictors include numerical variables like BMI (body mass index) and SleepTime (hours of sleep per day) and categorical variables like Smoking, AlcoholDrinking, Stroke, Sex, AgeCategory, and Race. Health-related features such as PhysicalHealth (days with poor physical health), MentalHealth (days with poor mental health), and DiffWalking (difficulty walking) are also included, along with pre-existing conditions like Diabetic, Asthma, KidneyDisease, and SkinCancer. PhysicalActivity and GenHealth reflect activity levels and self-rated general health. The large dataset is ideal for predictive modeling, exploring health patterns, and identifying key contributors to heart disease[51].

The Key Indicators of Heart Disease dataset was chosen for the study because it provides a large-scale, feature-rich foundation for predictive modeling and health risk assessment. The dataset is well-suited for big data processing using PySpark, enabling efficient handling of large volumes of health-related data. Additionally, its origin from the CDC's 2020 annual survey ensures data reliability and public health relevance, making it suitable for identifying high-risk populations and informing preventive healthcare strategies. While the dataset is rich in health-related features and public health relevance, a deeper exploration of biases, class imbalances, and self-reporting limitations is necessary to ensure accurate, ethical, and generalizable insights. Conducting data distribution analysis, fairness assessments, and bias mitigation strategies will enhance the dataset's utility for predictive modeling and healthcare decision-making.

4.2. Correlation and Feature Importance Analysis

The section defines a set of feature importance values based on individual feature weights for various health-related factors, such as BMI, PhysicalHealth, MentalHealth, Smoking, AlcoholDrinking, PhysicalActivity, and others. These weights are used to compute an individual's HI by calculating a weighted sum of the valid features present in the dataset. The correlation of each feature with the target variable (HeartDisease) is computed, and these correlations are normalized to obtain feature weights. Additionally, SHAP analysis and feature importance using XGBOOST and Random Forest algorithm are employed to calculate feature weights, see Table 2.

Table 2. Comparison of Feature Weight.

Features	Correlation	SHAP	XGBoost-F1	RF-FI
BMI	0.0318	0.0225	0.2383	0.0314
PhysicalHealth	0.1047	0.0671	0.0858	0.0516
MentalHealth	0.0175	0.0031	0.0946	0.0172
SleepTime	0.0051	0.0178	0.1039	0.0264
Smoking	0.0661	0.1596	0.0266	0.0173
AlcoholDrinking	0.0197	0.0129	0.0121	0.0028
Stroke	0.1207	0.0773	0.0252	0.1523
DiffWalking	0.1234	0.0939	0.0261	0.0986
AgeCategory	0.0817	0.6925	0.1194	0.1441
Diabetic	0.0765	0.6924	0.0419	0.0764
PhysicalActivity	0.0613	0	0.0208	0.0062
GenHealth	0.1029	0.3555	0.0661	0.239
Asthma	0.0254	0.0327	0.0237	0.0057
KidneyDisease	0.089	0.0433	0.0185	0.0519
SkinCancer	0.0572	0.0417	0.0145	0.0126

4.3. Comparison of Classification Results

Different feature importance methods—SHAP Analysis, XGBoost, Random Forest, and Correlation Analysis—demonstrate varying results. The accuracy, precision, recall, and specificity levels prove how well each method distinguishes between the two risk categories. A comparison of classification results is presented in Table 3.

Table 3. Comparison of predicting results based on varying Feature Weight.

Metrics	Correlation	SHAP	XGBoost-F1	RF-FI
Accuracy	0.9774	0.7276	0.6586	0.6676
Precision	1.0000	0.2132	0.1294	0.1466
F1 Score	0.9940	0.3377	0.2067	0.2347
Recall	0.9890	0.8116	0.5135	0.5886
MCC	0.9890	0.3171	0.1101	0.1557
Balanced Accuracy	0.9940	0.7656	0.5930	0.6318
Specificity	1.0000	0.7197	0.6724	0.6750
Optimal Threshold	1.0000	0.0825	2.1251	1.2181

SHAP Analysis prioritizes high recall (81.16%), ensuring most high-risk individuals are identified. However, its low precision (21.32%) suggests a high number of false positives, meaning many low-risk individuals may be misclassified as high-risk. The low optimal threshold (0.0825) indicates that even minor deviations in health indicators can push individuals toward a higher risk category, making SHAP a highly sensitive but less specific approach. XGBoost, on the other hand, has moderate recall (51.35%) and low precision (12.94%), meaning it is more conservative in assigning individuals to the high-risk category but may fail to identify some truly high-risk cases. Its higher threshold (2.1251) suggests stricter

classification criteria, reducing false positives but potentially misclassifying some high-risk individuals as low-risk. Random Forest performs slightly better than XGBoost, with a recall of 58.86% and precision of 14.66%, meaning it improves high-risk detection but still struggles to separate high-risk from low-risk individuals. Correlation Analysis, however, achieves perfect precision (100%) and very high recall (98.90%), ensuring that all individuals classified as high risk truly belong in that category. The optimal threshold (1.0000) establishes a strict decision boundary, ensuring clear separation between risk levels. Among the evaluated methods, Correlation Analysis outperforms all others in classifying individuals into low-risk and high-risk categories. With a high accuracy (97.74%), perfect precision (100%), high recall (98.90%), and optimal balanced accuracy (99.40%), it ensures that high-risk individuals are correctly identified while minimizing false positives.

Unlike SHAP, XGBoost, and Random Forest, which struggle with either low precision or recall, Correlation Analysis provides a well-defined separation between risk categories, leading to more reliable health assessments. Its superior performance in both sensitivity and specificity makes it the most effective method for accurate health risk classification. This demonstrates that a correlation-based approach is highly suitable for optimizing health predictions and ensuring precise, data-driven recommendations for personalized health interventions

4.4. Point-Biserial Correlation Between Health Index and Features

Understanding the relationship between health outcomes and various lifestyle or demographic factors is crucial for identifying potential risk factors and informing public health strategies. In this analysis, we explore the Point-Biserial Correlation between the HI and specific features to quantify the strength and significance of their association. Table 4 shows the Point-Biserial Correlation Between HI and the Features within the heart disease dataset used.

Table 4. Depicts the Explanation of the Feature Weights of the Heart Disease Dataset

Features	Point-Biserial Correlation
BMI	0.3050
Smoking	0.1376
AlcoholDrinking	-0.0034
Stroke	0.1488
DiffWalking	0.4512
AgeCategory	0.1977
Diabetic	0.1592
PhysicalActivity	0.2735
GenHealth	0.4962
Asthma	0.1613
KidneyDisease	0.1539
SkinCancer	0.0055
HeartDisease	0.1582
MentalHealth	0.4368
PhysicalHealth	0.9127

The key observations are based on Table 4 are

1. Strong Positive Correlations:
   - PhysicalHealth (0.9127): This feature has the strongest positive correlation with the HI, indicating that better physical health is highly associated with a higher HI.
   - GenHealth (0.4962) and MentalHealth (0.4368): These also show strong positive relationships, suggesting that general and mental health significantly contribute to the overall HI.
   - DiffWalking (0.4512): Difficulty walking is moderately associated with the HI, likely indicating that mobility issues negatively impact health.

2. Moderate Positive Correlations:
   - BMI (0.3050) and PhysicalActivity (0.2735): These features show moderate positive relationships, implying that higher BMI and physical activity levels are associated with better health outcomes.
   - Asthma (0.1613), KidneyDisease (0.1539), Stroke (0.1488), Diabetic (0.1592) and Smoking (0.1376): These conditions or behaviors have weaker but still notable positive correlations with the HI.
3. Weak or Negligible Correlations:
   - AlcoholDrinking (-0.0034): The negative value suggests a very weak inverse relationship, meaning alcohol consumption has almost no impact on the HI in this dataset.
   - SkinCancer (0.0055): This feature show almost no correlation with the HI, indicating that it has little to no influence on health outcomes in this context.
4. AgeCategory (0.1217): has a weak positive correlation, suggesting that older age groups may have slightly higher HI scores, though the relationship is not strong.

4.5. Personalized Health Recommendation

The result for assessing health risk and generating personalized health recommendations for the first ten records is shown in Table 5. The analysis aims to assess individual health risks and provide tailored recommendations by leveraging a dataset of personal health indicators. Using a combination of feature engineering, clustering techniques, and domain-specific weighting, the approach integrates multiple health-related factors, such as physical activity, mental health, chronic conditions, and lifestyle habits, into a single composite score called the HI. This index is designed to quantify overall health status, reflecting both positive and negative contributors. By categorizing individuals into risk groups through clustering, the analysis enables precise identification of those requiring lifestyle improvements or medical attention. The framework assesses health risks and offers actionable recommendations based on specific health behaviors and conditions. This structured methodology highlights the value of personalized, data-driven health risk evaluation and management approaches.

Table 5. Depicts Health Index-Based Risk Classification and Personalized Recommendations for Ten (10) records Using Correlation Analysis

No	Health Index	Risk Category	Personalized Recommendations
1	0.7324576311664053	Low Risk	Focus on improving physical health and managing stress. Incorporate regular physical exercise.
2	1.5017572619568902	High Risk	Seek professional medical advice and prioritize health improvement
3	5.270173077120121	High risk	Focus on improving physical health and managing stress. Incorporate regular physical exercise.
4	0.9466198197373792	Low Risk	Focus on improving physical health and managing stress. Incorporate regular physical exercise.
5	5.284683873891222	High Risk	Seek professional medical advice and prioritize health improvement.
6	1.065278147701063	High Risk	Seek professional medical advice and prioritize health improvement.
7	1.1229198466771444	High Risk	Seek professional medical advice and prioritize health improvement.
8	1.3728420189770663	High Risk	Seek professional medical advice and prioritize health improvement.
9	2.899728390139356	High Risk	Focus on improving physical health and managing stress. Incorporate regular physical exercise.
10	0.9695813185825066	Low Risk	Focus on improving physical health and managing stress. Incorporate regular physical exercise.

4.6. ROC and SHI Analysis

This section presents the evaluation ROC curve analysis, confusion matrix, and the distribution of SHI based optimal threshold selection using correlation feature weighting. These results provide insights into how well the model distinguishes between risk categories and the overall distribution of health risk scores. Figure 2 illustrates the ROC curve of the classification model using correlation-based feature weighting. The Area Under the Curve (AUC) = 0.994, indicating excellent discriminatory ability. The optimal threshold for classification is 1.000, as marked in red, which ensures an effective balance between sensitivity and specificity.

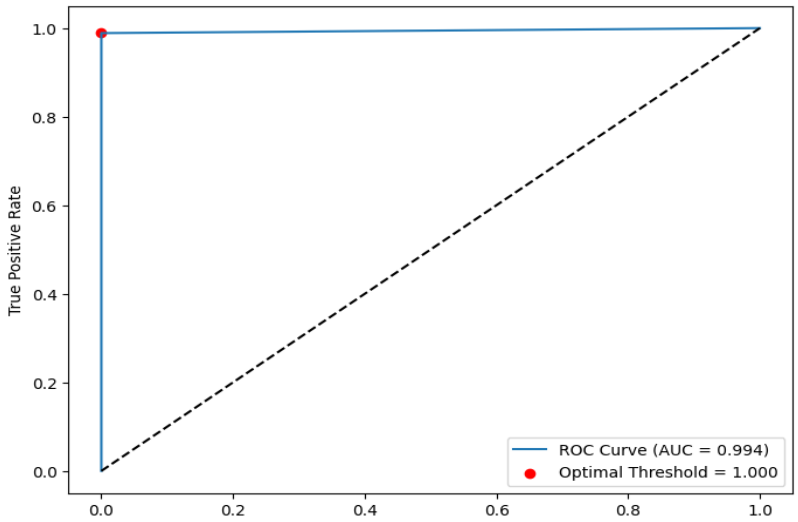

Figure 2. The Receiver Operating Characteristic (ROC) Curve with Optimal Threshold Selection Using Correlation Feature Weighting

Figure 3 displays the confusion matrix for the risk stratification model. The model correctly classifies 32,448 low-risk individuals with no false positives and correctly identifies 31,168 high-risk individuals. However, 355 high-risk cases were misclassified as low-risk, representing a small false-negative rate. This high classification accuracy further supports the effectiveness of correlation-based feature weighting in stratifying health risks.

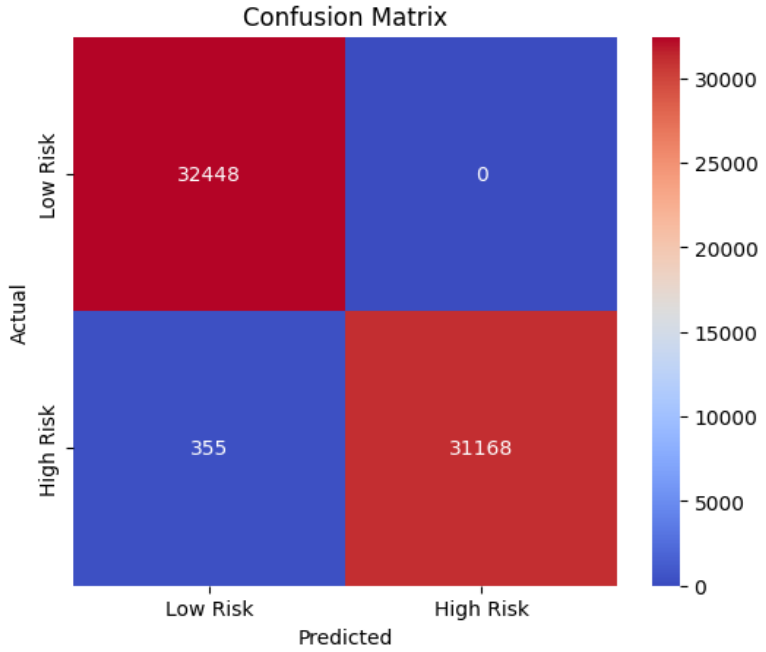

Figure 3. The Confusion Matrix for Risk Stratification Model Using Correlation Feature Weighting

Figure 4 presents the distribution of SHI scores among individuals in the dataset. The histogram follows a right-skewed distribution, with most individuals having SHI values between 1 and 2, indicating lower health risks. A secondary peak around SHI values between 4 and 5 suggests the presence of a distinct subgroup with higher health risks. The Kernel Density Estimation (KDE) curve highlights these trends, emphasizing that while most individuals have lower risk, a subset exhibits substantially higher SHI values, potentially linked to multiple chronic conditions or other risk factors. This analysis confirms that the HI effectively captures variations in health risk, allowing accurate classification and targeted health recommendations.

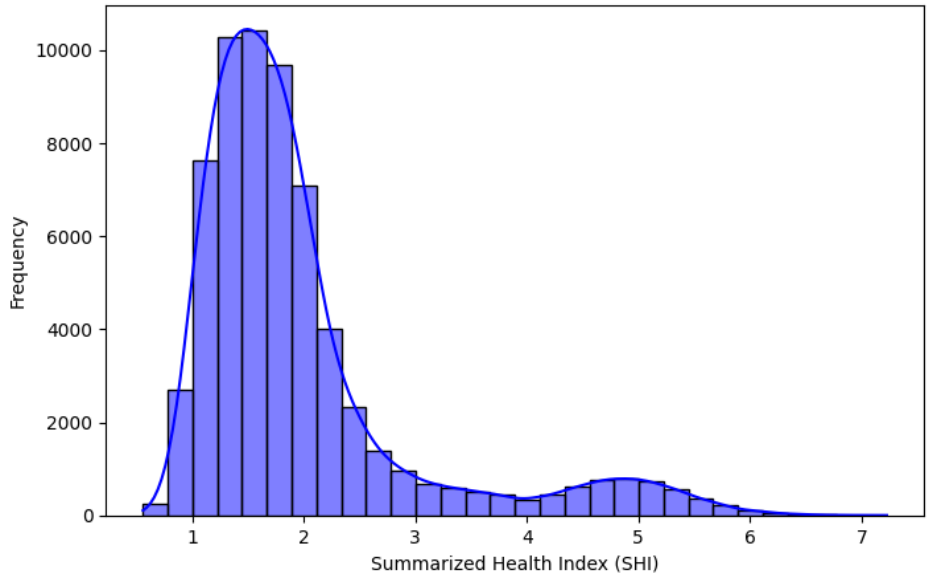

Figure 4. The Distribution of Summarized Health Index (SHI) Indicating Health Risk Variability

5. Conclusions

The proposed framework introduces a data-driven approach to health risk assessment and personalized recommendation generation, leveraging a combination of health indicators, clustering techniques, and domain-specific weighting. By computing HI that quantifies overall health status, the methodology effectively stratifies individuals into different risk categories and provides tailored recommendations to guide health management strategies.

The HI enables a personalized risk assessment, integrating multiple health-related factors to identify individuals requiring medical intervention or lifestyle modifications. Through data-driven decision-making, the approach facilitates targeted recommendations that enhance the precision of health interventions, ensuring that preventive measures and treatments are tailored to individual needs. Additionally, the methodology supports continuous health monitoring, allowing for longitudinal tracking of health trends and enabling adaptive risk management.

Beyond its application at the individual level, this framework has broader implications for healthcare optimization, offering a scalable solution for integrating machine learning into clinical decision support systems. Improving risk stratification and resource allocation contributes to more efficient healthcare delivery, minimizing unnecessary interventions while prioritizing high-risk individuals. This underscores the potential of Big Data analytics in shaping future healthcare strategies, bridging the gap between predictive modeling and actionable health insights.

Author Contributions: Conceptualization: O.A.A. and M.E.I.; methodology, O.A.A.; software: O.A.A.; validation: O.A.A., and M.E.I.; formal analysis: O.A.A.; investigation: M.E.I; resources: O.A.A.; writ-ing—original draft preparation: O.A.A.; writing—review and editing: M.E.I.; visualization: O.A.A; supervision: M.E.I; project administration: O.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this research is a well-known heart disease dataset available on the Kaggle platform at https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] R. Naqvi, T. R. Soomro, H. M. Alzoubi, T. M. Ghazal, and M. T. Alshurideh, "The Nexus Between Big Data and Decision-Making: A Study of Big Data Techniques and Technologies," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, Springer, Cham, 2021, pp. 838–853. doi: 10.1007/978-3-030-76346-6_73.

[2] C. Nyamful and R. Agrawal, "Big Variety Data," in *Encyclopedia of Big Data*, Cham: Springer International Publishing, 2022, pp. 110–113. doi: 10.1007/978-3-319-32010-6_23.

[3] A. T. Atieh, "The Next Generation Cloud technologies: A Review On Distributed Cloud, Fog And Edge Computing and Their Opportunities and Challenges," *Res. Rev. Sci. Technol.*, vol. 1, no. 1, pp. 1–15, 2021, [Online]. Available: https://researchberg.com/index.php/rrst/article/view/18

[4] S. Nazir *et al.*, "A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming," *IEEE Access*, vol. 8, pp. 95714–95733, 2020, doi: 10.1109/ACCESS.2020.2995572.

[5] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, p. 54, Dec. 2019, doi: 10.1186/s40537-019-0217-0.

[6] Z. M. Tun and M. Aye Khine, "Cardiac Diagnosis Classification Using Deep Learning Pipeline on Apache Spark," in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Jun. 2020, pp. 743–746. doi: 10.1109/ECTI-CON49241.2020.9158314.

[7] K. Batko and A. Ślęzak, "The use of Big Data Analytics in healthcare," *J. Big Data*, vol. 9, no. 1, p. 3, Dec. 2022, doi: 10.1186/s40537-021-00553-4.

[8] P. Kangelani and T. Iyamu, "A Model for Evaluating Big Data Analytics Tools for Organisation Purposes," in *Responsible Design, Implementation, and Use of ICT (Information and Communication Technology)*, 2020.

[9] D. Otoo-Arthur and T. L. van Zyl, "A Scalable Heterogeneous Big Data Framework for e-Learning Systems," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Aug. 2020, pp. 1–15. doi: 10.1109/icABCD49160.2020.9183863.

[10] R. Venkatraman and S. Venkatraman, "Big Data Infrastructure, Data Visualisation and Challenges," in *Proceedings of the 3rd International Conference on Big Data and Internet of Things*, Aug. 2019, pp. 13–17. doi: 10.1145/3361758.3361768.

[11] R. Rossi and K. Hirama, "Characterizing Big Data Management," *Issues Informing Sci. Inf. Technol.*, vol. 12, pp. 165–180, 2015, doi: 10.28945/2204.

[12] S. Acharjee and R. Choudhury, "Big data searching using words," *arXiv*. Sep. 10, 2024. [Online]. Available: http://arxiv.org/abs/2409.15346

[13] J. Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *J. Evid. Based. Med.*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: 10.1111/jebm.12373.

[14] S. Venkatraman and R. Venkatraman, "Big data security challenges and strategies," *AIMS Math.*, vol. 4, no. 3, pp. 860–879, 2019, doi: 10.3934/math.2019.3.860.

[15] S. Usman, R. Mehmood, I. Katib, and A. Albeshri, "Data Locality in High Performance Computing, Big Data, and Converged Systems: An Analysis of the Cutting Edge and a Future System Architecture," *Electronics*, vol. 12, no. 1, p. 53, Dec. 2022, doi: 10.3390/electronics12010053.

[16] S. Dasari and R. Kaluri, "Big Data Analytics, Processing Models, Taxonomy of Tools, V's, and Challenges: State-of-Art Review and Future Implications," *Wirel. Commun. Mob. Comput.*, vol. 2023, pp. 1–14, May 2023, doi: 10.1155/2023/3976302.

[17] A. Shanbhag, S. Madden, and X. Yu, "A Study of the Fundamental Performance Characteristics of GPUs and CPUs for Database Analytics," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, Jun. 2020, pp. 1617–1632. doi: 10.1145/3318464.3380595.

[18] E. Shaikh, I. Mohiuddin, Y. Alufaisan, and I. Nahvi, "Apache Spark: A Big Data Processing Engine," in *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, Nov. 2019, pp. 1–6. doi: 10.1109/MENACOMM46666.2019.8988541.

[19] M. Saxena, S. Jha, S. Khan, J. Rodgers, P. Lindner, and E. Gabriel, "Comparison of MPI and Spark for Data Science Applications," in *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, May 2020, pp. 682–690. doi: 10.1109/IPDPSW50202.2020.00123.

[20] M. Alam Mallik, N. Fariza Zulkurnain, S. Siddiqui, and R. Sarkar, "The Parallel Fuzzy C-Median Clustering Algorithm Using Spark for the Big Data," *IEEE Access*, vol. 12, pp. 151785–151804, 2024, doi: 10.1109/ACCESS.2024.3463712.

[21] S. Tang, B. He, C. Yu, Y. Li, and K. Li, "A Survey on Spark Ecosystem: Big Data Processing Infrastructure, Machine Learning, and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 1–1, 2020, doi: 10.1109/TKDE.2020.2975652.

[22] M. Gecer, "Debugging Spark Applications A Study on Debugging Techniques of Spark Developers Master Thesis," Universit ̈at Bern, 2020. [Online]. Available: https://scg.unibe.ch/archive/masters/Gece20a.pdf

[23] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," in *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, Apr. 2019, pp. 1–5. doi: 10.1109/WITS.2019.8723839.

[24] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, vol. 6, no. 1, p. 81, Dec. 2019, doi: 10.1186/s40537-019-0244-x.
[25] S. Ilbeigipour, A. Albadvi, and E. Akhondzadeh Noughabi, "Real-Time Heart Arrhythmia Detection Using Apache Spark Structured Streaming," *J. Healthc. Eng.*, vol. 2021, pp. 1–13, Apr. 2021, doi: 10.1155/2021/6624829.
[26] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning," *Appl. Sci.*, vol. 10, no. 4, p. 1398, Feb. 2020, doi: 10.3390/app10041398.
[27] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 714–722, Oct. 2020, doi: 10.1016/j.future.2019.09.056.
[28] A. Ed-daoudy, K. Maalmi, and A. El Ouaazizi, "A scalable and real-time system for disease prediction using big data processing," *Multimed. Tools Appl.*, vol. 82, no. 20, pp. 30405–30434, Aug. 2023, doi: 10.1007/s11042-023-14562-3.
[29] P. Rajendra Kumar, P. Chakrabarti, T. Chakrabarti, B. Unhelkar, and M. Margala, "Heart disease prediction using spark architecture with fused feature set and hybrid Squeezenet-Linknet model," *Biomed. Signal Process. Control*, vol. 100, p. 107070, Feb. 2025, doi: 10.1016/j.bspc.2024.107070.
[30] Y. K. Gupta and S. Kumari, "Performance Evaluation of Distributed Machine Learning for Cardiovascular Disease Prediction in Spark," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Jun. 2021, pp. 1506–1512. doi: 10.1109/ICOEI51242.2021.9452955.
[31] Arif Ahmad Shehloo and Ganesh Gopal Varshney, "Realizing the Potential of Big Data Analytics through Apache Spark MLlib," *Nanotechnol. Perceptions*, pp. 1813–1830, Nov. 2024, doi: 10.62441/nano-ntp.vi.3022.
[32] S. Eeti, "Real-Time Data Processing: An Analysis of PySpark's Capabilities," *Int. J. Res. Anal. Rev.*, vol. 8, no. 3, 2021, [Online]. Available: www.ijrar.org
[33] E. Dorison, F. Lesur, D. Meurice, and G. Roinel, "Health index, a tool for asset management," in *International Conference on Power Insulated Cables*, 2007. [Online]. Available: https://www.jicable.org/2007/Actes/Session_B4/JIC07_B41.pdf
[34] D. Kornbrot, "Point Biserial Correlation," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. doi: 10.1002/9781118445112.stat06227.
[35] J. D. Brown, "Point - biserial correlation coefficientsbiserial correlation coefficients," *Shiken: JLT Testing & Evlution SIG Newsletter*. pp. 13–17, 2001. [Online]. Available: https://teval.jalt.org/test/PDF/Brown12.pdf
[36] K. Pytlak, "Indicators of Heart Disease (2022 UPDATE)." 2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data
[37] I. Malakar and B. Nepal, "Conceptualizing Explorative Data Analysis in Applied Statistics," *Patan Gyansagar*, vol. 6, no. 1, pp. 46–63, Jul. 2024, doi: 10.3126/pg.v6i1.67406.
[38] M. Abt, T. Leuders, K. Loibl, and F. Reinhold, "Developing initial notions of variability when learning about box plots," *Math. Think. Learn.*, pp. 1–24, Oct. 2024, doi: 10.1080/10986065.2024.2421412.
[39] R. L. Nuzzo, "The Box Plots Alternative for Visualizing Quantitative Data," *PM&R*, vol. 8, no. 3, pp. 268–272, Mar. 2016, doi: 10.1016/j.pmrj.2016.02.001.
[40] J. H. Kwak, H. Bin Lee, and K.-H. Lee, "Exploring how to Organize a Unit on Box Plots Through Analysis of Foreign Textbooks," *Korean Soc. Educ. Stud. Math. - Sch. Math.*, vol. 25, no. 2, pp. 249–276, Jun. 2023, doi: 10.57090/sm.2023.06.25.2.249.
[41] K. Hu, "Become Competent within One Day in Generating Boxplots and Violin Plots for a Novice without Prior R Experience," *Methods Protoc.*, vol. 3, no. 4, p. 64, Sep. 2020, doi: 10.3390/mps3040064.
[42] E. Soltanmohammadi and N. Hikmet, "Optimizing Healthcare Big Data Processing with Containerized PySpark and Parallel Computing: A Study on ETL Pipeline Efficiency," *J. Data Anal. Inf. Process.*, vol. 12, no. 04, pp. 544–565, 2024, doi: 10.4236/jdaip.2024.124029.
[43] A. Senbato, "Designing Healthcare Data Analytics Framework Based on Big Data Approach: In Case of Stroke Disease Prediction," Addis Ababa Science and Technology University, 2019.
[44] K. Sharma *et al.*, "Apache Spark for Analysis of Electronic Health Records: A Case Study of Diabetes Management," *Rev. d'Intelligence Artif.*, vol. 37, no. 6, pp. 1521–1526, Dec. 2023, doi: 10.18280/ria.370616.
[45] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. London, England: MIT Press, 2016.
[46] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?,'" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Augu, pp. 1135–1144. doi: 10.1145/2939672.2939778.
[47] C. M. Bishop and N. M. Nasrabadi, "Pattern Recognition and Machine Learning," *J. Electron. Imaging*, vol. 16, no. 4, p. 049901, Jan. 2007, doi: 10.1117/1.2819119.
[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd editio. Springer, 2017.
[49] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006, doi: 10.1016/j.ymssp.2005.09.012.
[50] G. Niu, T. Han, B.-S. Yang, and A. C. C. Tan, "Multi-agent decision fusion for motor fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 3, pp. 1285–1299, Apr. 2007, doi: 10.1016/j.ymssp.2006.03.003.
[51] M. J. Goddard and I. Hinberg, "Receiver operator characteristic (ROC) curves and non-normal data: An empirical study," *Stat. Med.*, vol. 9, no. 3, pp. 325–337, Mar. 1990, doi: 10.1002/sim.4780090315.