5th International Conference on Electronics and Sustainable Communication Systems (ICESC 2024)
IEEE XPlore Part Number: CFP24V66-ART; ISBN: 979-8-3503-7994-5

Cardiovascular Disease Prediction using Machine Learning-Random Forest Technique

Malarkodi P[1], Arun M[2], Manikandan R[3], Ramkumar S[4]
[1,2]*Department of Computer Applications*
[1,2]*Kalasalingam Academy of Research and Education*
Krishnankoil, Tamilnadu, India
[3]*School of Business and Management,* [4]*School of Sciences,*
[3,4]*Christ University,*
Bengaluru, Karnataka, India
malarkodi79p@gmail.com[1], vgsm.arun@gmail.com[2], [3]manikandan.rajagopal@christuniversity.in, [4]ramkumar.s@christuniversity.in
Corresponding Author: vgsm.arun@gmail.com

2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-7994-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICESC60852.2024.10689896

***Abstract*—Cardiovascular diseases (CVDs) pose a significant global health challenge. Early and accurate diagnosis is crucial for effective treatment. This research focuses on developing a robust classification system for CVDs using machine learning techniques. This study proposes an enhanced Random Forest (RF) model optimized for big data environments and explore the potential of CNN-based classification. By leveraging medical imaging data and employing these advanced algorithms, we aim to improve the accuracy and efficiency of CVD diagnosis.**

Keywords—Cardiac Disease Analysis, Convolution Neural Network (CNN), Machine Learning-Random Forest (ML-RF), Deep Learning Algorithm, Big Data Architecture, Image Processing.

I. INTRODUCTION

Cardiovascular disease (CVD) poses a significant challenge in clinical data analysis, driving researchers to explore predictive methods using machine learning (ML) techniques. This focus is driven by the high prevalence of heart-related disorders, occurred due to modern lifestyle changes such as increased stress, poor diet, and sedentary behavior, which collectively contribute to a rising incidence of cardiovascular diseases worldwide [1]. Machine learning techniques offer an innovative approach to predicting cardiovascular diseases by analyzing and interpreting complex medical data, often outperforming traditional statistical methods. These techniques are particularly adept at handling the intricate interactions between genetic, environmental, and lifestyle factors that contribute to heart disorders, which range from heart failure to coronary artery disease [2]. The integration of big data architecture with cardiovascular medicine holds tremendous potential for advancing precision healthcare. Big data analytics, supported by scalable infrastructure and machine learning algorithms, allows for the handling, examination, and interpretation of vast and heterogeneous datasets. By combining data from multiple sources—such as genetic information, wearable technology, electronic health records, and lifestyle data—researchers and medical professionals can achieve a comprehensive and dynamic understanding of an individual's heart health profile. This approach facilitates data-driven, personalized interventions that could revolutionize cardiovascular care.

This article is organized as follows: Section 2 covers related work in the field, while Section 3 provides a block diagram and detailed algorithmic descriptions. Section 4 explains the proposed method, including experiments and the prediction system, and compares and validates the outcomes of the proposed ML-AJ model using various parameters. Finally, Section 5 concludes with the findings and implications of the research.

II. RELATED WORKS

The work by [3] investigates the application of deep learning for cardiac image segmentation, highlighting its ability to enhance the precision of biomedical image analysis. The study also discusses the limitations of current deep learning methods, such as the lack of labeled data, challenges in interpretability, and issues related to the generalizability of models across different domains. It suggests future research directions to address these challenges. Reference [4] reviews the technological progress in healthcare, particularly in cardiac services. It provides an in-depth analysis of various machine learning methods used to predict cardiovascular events, discussing the strengths and weaknesses of each. The study emphasizes that five proposed models outperform previous iterations, justifying the use of AI in cardiology. The study in [5] explores the challenges and fundamental concepts involved in managing large biological datasets in healthcare. It identifies five key areas where big data applications can be seen, including public health awareness, healthcare ecosystem stakeholder interactions, hospital management, specific medical conditions, and healthcare technology delivery. The systematic literature review (SLR) offers valuable insights for both theoretical and practical applications, as well as future research directions.

The research in [6] discusses the application of machine learning techniques in automated image processing for disease diagnosis and detection. The study covers a broad range of data-related challenges, from simple analytical queries to the complex issues of analyzing raw images. It highlights the dual role of machine learning in generating therapeutically significant knowledge and automating tasks that would otherwise be performed by humans. The research work referenced in [7] evaluates various supervised learning algorithms, such as Decision Trees (DT), Random Forest (RF), Naïve Bayes, and K-Nearest Neighbour (KNN), in the context of medical data analysis. It examines the effectiveness of ensemble learning methods in predicting cardiovascular diseases, providing insights into their

predictive power. In [8], the focus is on using convolutional neural networks (CNNs) for biomedical image classification. The paper reviews different CNN architectures and their performance in categorizing medical images, discussing the evolution of CNNs and their impact on disease prediction. The study in [9] investigates the use of machine learning techniques to identify key features in predicting cardiovascular diseases. It combines various feature sets with well-known classification techniques to create a high-performance heart disease prediction model. The study also compares significant factors contributing to cardiovascular disease, using the Internet of Medical Things (IoMT) platform.

Reference [10] explores how big data architecture can be integrated into cardiovascular disease prediction models. The paper highlights the potential for improved scalability and accuracy through advanced information processing technology and database management techniques. The research in [11] highlights the role of machine learning, particularly deep learning, in advancing medical image analysis. It contrasts traditional hand-crafted features with the autonomous feature learning capabilities of neural networks, highlighting the significant progress in biomedical engineering. Finally, [12] discusses the application of AI and machine learning in healthcare, with a specific focus on peripheral artery disease. The paper provides an overview of key concepts and examines the current state of these technologies in diagnosing and predicting disease, suggesting areas where advanced analytics could further improve care.

III. PROPOSED METHODOLOGY

The proposed system utilizes a training set of observations to perform classification tasks effectively. This system integrates a preprocessing phase within the data layers that constitute the large image, which is crucial for handling and processing vast amounts of data. The preprocessing phase is essential as it involves extracting features from the large-scale images and storing this temporary data using big data architecture. This approach ensures that the data is efficiently managed and available for further analysis.

In this study, an enhanced machine learning technique called the Random Forest (ML-RF) is proposed, which can be applied within a big data architecture framework. The ML-RF technique is specifically designed to handle the complexities and scale of big data, making it suitable for applications that involve large image datasets. This study also suggests an organizational methodology based on the identification of the best-performing algorithm. The existing Convolutional Neural Network (CNN) method is noted for its varying performance in image storage classification, which highlights the need for an improved approach.

The ML-RF technique is introduced as a solution to these challenges. It employs big data to store temporary data after the large-scale images have undergone preprocessing and feature extraction. In Fig. 1, the Random Forest approach is illustrated, showcasing its use with different classification tools to enhance the overall process and improve the performance of the system's outcomes. This method ultimately boosts the effectiveness of machine learning in data storage and processing, particularly in the context of handling large image datasets within a big data environment.

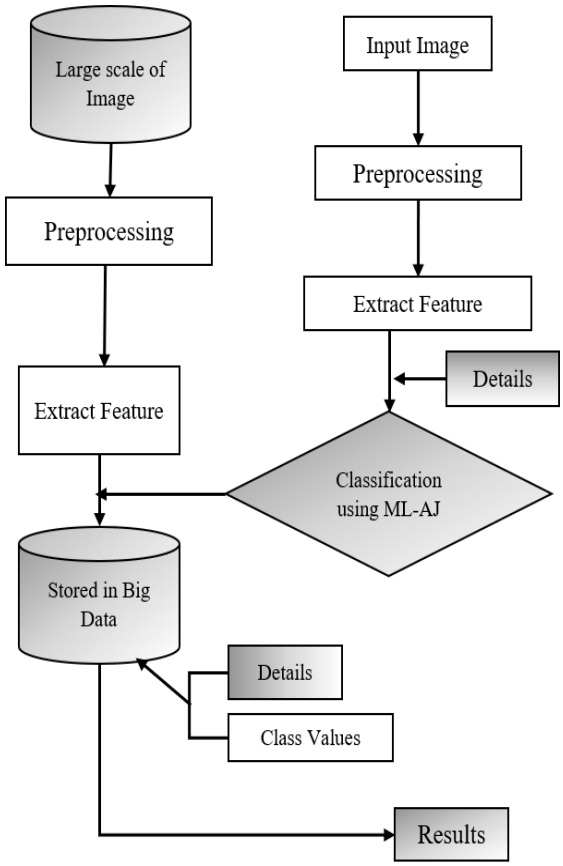

Figure.1. Proposed Methodology

The implementation of the ML-RF (Machine Learning-Random Forest) technique aims to significantly enhance machine learning capabilities within the field of big data, particularly for biomedical image analysis in cardiovascular disease prediction. This technique focuses on improving classification accuracy by combining advanced image processing techniques with the robust classification power of Random Forest algorithms. By utilizing these methods, ML-RF enhances data collection accuracy and the quality of the extracted features, leading to more precise and reliable predictions.

A. Data Collection

To ensure that the model can effectively analyze and predict cardiovascular diseases, it is crucial to gather a diverse dataset of biomedical images that represent various cardiovascular conditions. These images should come from multiple imaging modalities and cover a wide spectrum of disease types. This diversity in image attributes helps the model generalize better, making it more robust and accurate in its predictions.

SHAP (Shapley Additive explanations) features are employed to enhance the interpretability and transparency of the machine learning model. By using SHAP, we can better understand the contributions of individual features to the model's predictions, which is particularly important in a healthcare setting where explainability is crucial. Hadoop, as the big data platform, is integral to this approach. Its distributed computing capabilities allow for the efficient handling, processing, and storage of large datasets, making it well-suited for managing the extensive biomedical imaging data required for this study.

The ML-RF technique is implemented within the Hadoop framework to leverage its capabilities in handling large-scale data. Hadoop's architecture supports the parallel processing of large datasets, which is essential for efficiently managing the voluminous biomedical images involved in cardiovascular disease prediction. This integration ensures that the system can scale as needed while maintaining performance. One of the key challenges in big data processing is the ability to perform real-time analysis. The Lambda architecture addresses this by combining real-time data processing with batch processing. In the ML-RF approach, this is achieved by first processing data in a batch layer to create a comprehensive view, and then integrating real-time data processing to update this view as new data arrives. This hybrid approach ensures that the model remains up-to-date with the latest data, improving its accuracy and responsiveness. Various technologies and methodologies are employed to optimize the performance of the ML-RF system:

Amazon Web Services (AWS) and Microsoft Azure:

These platforms provide robust tools for integrating CNN architecture-based batch and real-time network processing, enhancing the system's ability to handle large-scale data analysis.

Apache Hadoop and Apache Spark:

These tools are crucial for implementing the ML-RF technique, as they provide the necessary infrastructure for distributed computing and real-time data processing. The combination of Hadoop's YARN resource management and Spark's real-time processing capabilities ensures that the system can efficiently manage and analyze large datasets.

NoSQL Databases:

In the preprocessing layer of Hadoop, NoSQL databases are utilized to manage the large volumes of data generated by the biomedical images. This ensures that the data is stored efficiently and can be accessed quickly for analysis.

The current cardiovascular disease prediction model considers factors related to lifestyle like drinking, smoking, eating a poor diet, exercising, and managing stress. However, new research indicates that a wide range of other factors, including health conditions, social and economic factors, atmospheric environment information, and climate change, also have an impact on cardiovascular disease. However, it is not possible to extract risk factors that exist in numerous locations using an Integration database (DB) based on distinct factors related to the environment. Furthermore, there isn't a single best forecasting system that takes into account all of these variables. In order to provide a systematic health care strategy, it is vital to combine the environmental elements that have been found to affect health, such as unusual weather and pollution in the atmosphere, into numerical evaluations and diagnoses of health. The ML-RF technique represents a significant advancement in the application of machine learning and big data architecture to cardiovascular disease prediction. By leveraging the power of Random Forest algorithms, image processing, and big data tools like Hadoop, this approach offers improved accuracy and scalability, making it a valuable tool in the ongoing efforts to enhance healthcare outcomes.

B. Hadoop and Shap Features for Model Interpretability (ML-AJ Algorithm)

In some situations, like managing big datasets effectively and learning about the model's decision-making procedures, the combination of the two can be helpful.We can drastically cut down on the amount of time needed for image processing in the medical industry by utilizing Hadoop's distributed computing capabilities[25]. A method for equitably allocating each feature's contribution to the prediction for a specific instance in a predictive model is to use SHAP values. This is especially helpful for comprehending how each feature affects the predictions made by the model which was shown in the equ.1.

$$\text{Shap Value (f)} = \emptyset_i(f) = \sum_{s \subseteq N \setminus \{i\}} \frac{|s|!(|n|-|s|-1)!}{|N|!} [f(S \cup \{i\}) - f(s)] \quad (1)$$

Where,

- $\emptyset_i(f)$ is the Shap value for feature i in model f.
- n is a set of all factors.
- s is the subset of features (excluding i)
- F(s) is the model prediction with features in subset S.
- F(S∪ {i}) is the model prediction with features in subset S and i included.

From the equation (1) the integration of Hadoop's efficient distributed processing capabilities with SHAP features for model interpretability, it is possible to enhance the precision and lucidity of machine learning models, hence facilitating more informed decision-making.

C. An Improved Big Data Architecture Cardiovascular Disease Prediction System

Three main components can build the strong big data architecture on which the suggested prediction system is based: real-time analytics, predictive modelling, and data collecting and preparation. These components work together to process and analyzed data, build predictive models, and provide real-time insights and alerts for personalized healthcare. The system efficiently captures the complexity of cardiovascular disease risk by utilizing machine learning

methods including Support Vector Machines, Random Forests, and Deep Learning Models. This allows for accurate predictions and timely interventions. Overall, the proposed prediction system utilizes advanced computational approaches, data mining methods, and machine learning algorithms to develop accurate prediction models with high-performance capacity. Fig. 2 represents the level of accuracy achieved by the prediction system and its ability to generate real-time predictive alerts over a short period of time. The suggested graph is the level of prediction based on the interpretation and accuracy of machine learning algorithms (i.e., ML-AJ).

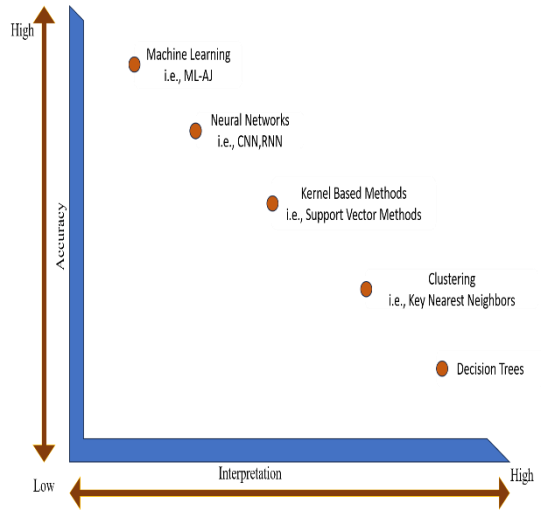

Figure 2. Data Accuracy and Interpretation

The proposed prediction system integrates real-time analytics, predictive modelling, and data collection and preparation to provide personalized healthcare through timely interventions. It leverages machine learning algorithms to accurately predict cardiovascular risk and generate real-time predictive alerts. These advancements in machine learning have the potential to revolutionize cardiovascular medicine by enabling proactive and targeted interferences, ultimately leading to better patient outcomes and overall healthcare service quality[26]. The integration of machine learning algorithms in cardiovascular risk prediction has shown promising results in improving accuracy and providing personalized healthcare.

IV. RESULTS & DISCUSSION

The example data utilised in the suggested system is displayed in Table 1. There are several file and data kinds included in these example data. Pre-processing different kinds of files might be costly. The system suggested in this study uses a data integration feature of the batch layer to process previously gathered file data and a real-time data integration feature of the speed layer to handle real-time data. Each image processes data in a different format, but the task is the same for all of them. We assessed MapReduce's performance in Hadoop batch layer with the highest data preparation cost in order to gauge the effectiveness of the suggested approach.

Table 1. Information from the suggested approach

DATA	TYPE	SIZE	PROVIDER
Statistics about health insurance	API File (XLS)	2001-2017 10 files (7MB)	Healthcare Bigdata Hub
Core National Open Data	API (JSON, XML) File (CSV)	1998-2017 10 Files (7 MB)	National Information Society Agency
Survey of Nutrition and health in the nation	File (SAV)	1998 -2015 60files (7 GB)	Korea Centres for Disease control & Prevention
Climate Data	API File (XML)	1999-2017 10 Files(7 MB)	Korea Meteorological Administration
Cardiac MRI Dataset	MRI IMAGE	2022 (3 GB)	Kaggle

Because the suggested approach integration classified image MapReduce jobs are set up as map-side only jobs, block-by-block parallel processing is feasible. By doing this, Reduce Jobs are removed, preventing performance degradation while collecting data. Fig. 3 depicts a straightforward an attempt at creating a map rather than designing the work as a map-side-only task, reduce it to a MapReduce task.

The distinction is that MapReduce tasks only contain the mapper's code; The dummy reduction produces an intermediate result with a null key value, but it does nothing else work. To find variations in performance based on the kind of MapReduce task, the trials were run in stand-alone mode. As illustrated in Fig. 3 these data performance raises in time with significant research with no time raised 200% in comparison to the task given with the details of data raises. Support vector machines, random forests, and deep learning models are just a few of the machine learning techniques that the system uses to effectively represent the complexity of cardiovascular disease risk.

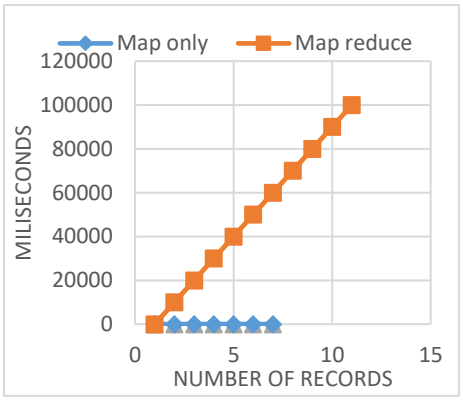

Figure 3. Performance of execution time using ML-AJ with Map Reduce in Hadoop.

V. RESULTS

The proposed approach was simulated and its effectiveness has been assessed using several metrics. The efficacy of the approaches in predicting diseases is evaluated based on many features and their respective values. Our evaluation's outcome was examined using AJ-ML method to compare the effectiveness of various approaches with those that are currently in use. Result of our study were explained in the Table.2 and Table.3

Table 2. Performance Analysis of Various Strategies

Model	KNN	RF	SVM	LR	SGD	MLP	Proposed
Precision	0.84	0.91	0.78	0.61	0.63	0.1	0.91
Recall	0.90	0.75	0.79	0.61	0.52	0.01	0.91
F-measure	0.87	0.83	0.79	0.61	0.57	0.02	0.91

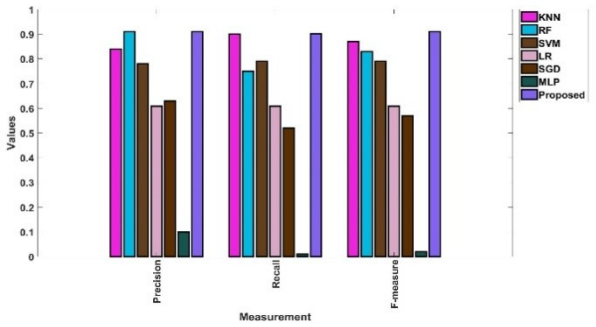

Figure 4. Performance Comparison of various machine learning algorithm proposed system

TABLE 3. Displays the accuracy of segmentation huge data for disease prediction; compared to other methods, the proposed ML-strategy achieved higher classification accuracy.

TABLE 3: PERFORMANCE COMPARISON

Model	KNN	RF	SVM	LR	SGD	MLP	Proposed
Accuracy	87	82.1	81.5	75.5	56.5	45.7	91.2

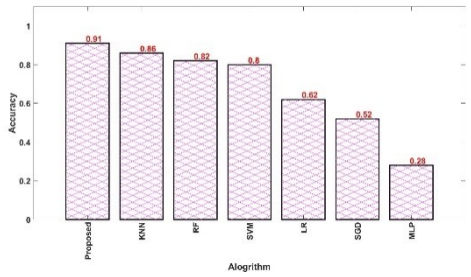

Figure 5. Current disease accuracy

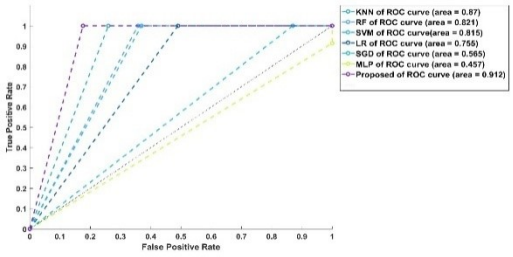

Figure 6. Graph Indicates true and false positive rate

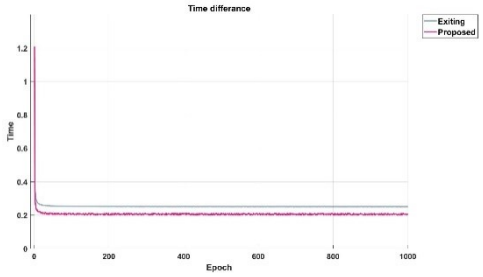

Figure 7. This graph indicates the time difference of the sample

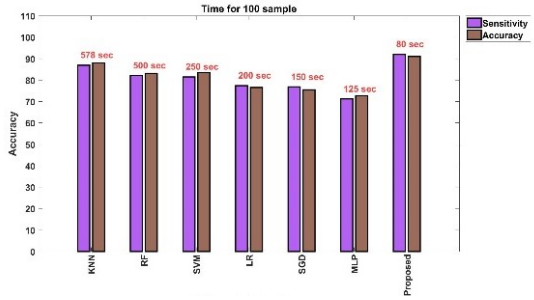

Figure 8. Sample and their time with sensitivity/accuracy

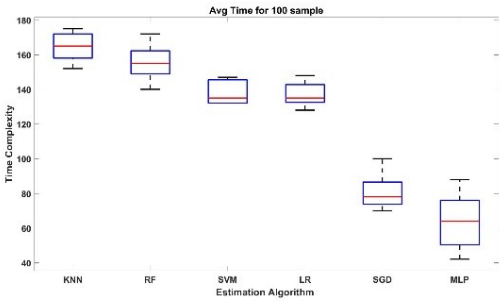

Figure 9. Sample test time estimation

Finally, the accuracy of disease prediction from Fig. 4 to Fig. 9 generated by various approaches as these chart level are as measured. Compared to other approaches in each class, our proposed method AJ-ML strategy has generated higher illness prediction.

VI. CONCLUSION

The integration of big data and machine learning technologies holds significant potential for transforming cardiovascular disease (CVD) prediction and prevention. By harnessing the power of these technologies, researchers can analyze vast amounts of patient data to identify patterns, trends, and risk factors associated with CVD. The use of machine learning algorithms, such as random forests, support vector machines, and neural networks, has shown promising results in predicting the likelihood of developing CVD. These models can effectively analyze complex medical data, including electronic health records, genetic information, and wearable device data, to identify individuals at high risk. To fully realize the potential of machine learning in CVD prediction, robust data infrastructure is essential. The adoption of big data architectures, such as Hadoop, enables efficient data storage, processing, and analysis. This facilitates the development and deployment of scalable machine learning models.

REFERENCES

[1] D. Roy, Md. A. Mahmood, and T. Joyti Roy, "An Analytical Model for Prediction of Heart Disease using Machine Learning Classifiers." Jun. 30, 2021. doi: 10.36227/techrxiv.14867175.v1.

[2] M. Shehab *et al.*, "Machine Learning in Medical Applications: A review of state-of-the-art methods".

[3] "Chen et al. - 2020 - Deep Learning for Cardiac Image Segmentation A Re.pdf."

[4] "Moshawrab et al. - 2023 - Predicting Cardiovascular Events with Machine Lear.pdf."

[5] S. Khanra, A. Dhir, A. K. M. N. Islam, and M. Mäntymäki, "Big data analytics in healthcare: a systematic literature review," *Enterp. Inf. Syst.*, vol. 14, no. 7, pp. 878–912, Aug. 2020, doi: 10.1080/17517575.2020.1812005.

[6] M. Henglin, G. Stein, P. V. Hushcha, J. Snoek, A. B. Wiltschko, and S. Cheng, "Machine Learning Approaches in Cardiovascular Imaging," *Circ. Cardiovasc. Imaging*, vol. 10, no. 10, p. e005614, Oct. 2017, doi: 10.1161/CIRCIMAGING.117.005614.

[7] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques : a survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[8] M. Bharath Simha Reddy and P. Rana, "Biomedical image classification using deep convolutional neural networks – overview," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012020, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012020.

[9] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," *IEEE Access*, vol. 8, pp. 59247–59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

[10] 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.

[11] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," *J. Med. Syst.*, vol. 42, no. 11, p. 226, Nov. 2018, doi: 10.1007/s10916-018-1088-1.

[12] A. M. Flores, F. Demsas, N. J. Leeper, and E. G. Ross, "Leveraging Machine Learning and Artificial Intelligence to Improve Peripheral Artery Disease Detection, Treatment, and Outcomes," *Circ. Res.*, vol. 128, no. 12, pp. 1833–1850, Jun. 2021, doi: 10.1161/CIRCRESAHA.121.318224.

[13] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing".

[14] B. Gupta and D. K. Jyoti, "Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data," vol. 5, 2014.

[15] B. Familiar and J. Barnes, "Business in Real-Time Using Azure IoT and Cortana Intelligence Suite".

[16] O.-C. Marcu, R. Tudoran, B. Nicolae, A. Costan, G. Antoniu, and M. S. Pérez-Hernández, "Exploring Shared State in Key-Value Store for Window-Based Multi-Pattern Streaming Analytics".

[17] S. H. Han, K. O. Kim, E. J. Cha, K. A. Kim, and H. S. Shon, "System Framework for Cardiovascular Disease Prediction Based on Big Data Technology," 2017.

[18] K. Sideris, R. Nejabati, and D. Simeonidou, "Seer: Empowering Software Defined Networking with Data Analytics".

[19] S. Dolev, P. Florissi, E. Gudes, S. Sharma, and I. Singer, "A Survey on Geographically Distributed Big-Data Processing using MapReduce," Jul. 2017. doi: 10.1109/TBDATA.2017.2723473.

[20] "performance comparision.pdf."

[21] N. Bagwari and O. Kumar, "Indexing optimizations on Hadoop," 2017.

[22] "integrating sql.pdf."

[23] R. Kumar, B. B. Parashar, S. Gupta, Y. Sharma, and N. Gupta, "Apache Hadoop, NoSQL and NewSQL Solutions of Big Data.," vol. 1, no. 6.

[24] C. Francalanci, I. P. Ravanelli, T. di L. di, and A. Decaneto, "Design and testing of an active Big Data architecture for social and crowding Emergency Management".

[25] A. Rehman, S. Naz, and I. Razzak, "Leveraging Big Data Analytics in Healthcare Enhancement: Trends, Challenges and Opportunities." arXiv, Apr. 05, 2020. Accessed: Jun. 17, 2024. [Online]. Available: http://arxiv.org/abs/2004.09010

[26] R. O'Shea, A. S. Ma, R. V. Jamieson, and N. M. Rankin, "Precision medicine in Australia: now is the time to get it right," *Med. J. Aust.*, vol. 217, no. 11, pp. 559–563, Dec. 2022, doi: 10.5694/mja2.51777.