IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predictive Analytics Of Cardiovascular Disease Using LSTM Networks

[1]Ms. Pritibala Sudhakar Ingle*, [2]Dr. Santosh Deshpande

[1]PhD Scholar, Computer Management, Institute of Management & Career Courses (IMCC), Research Centre, Pune, SPPU, India

[2]PhD Guide, Computer Management, Institute of Management & Career Courses (IMCC), Research Centre, Pune, SPPU, India

Abstract

Cardiovascular disease (CVD) is recognized as a major contributing factor to total global mortality, causing approximately 18 million deaths around the globe each year. It is consequently imperative that diagnostics be extremely reliable. Although machine learning, as well as ensemble-based approaches, have also exhibited considerable potential for clinical risk assessment, these have generally been seen to fail in addressing complex, nonlinear relationships typically seen in large clinical data sets. Leveraging a series of works that have created a new baseline of 92% tested ensemble approaches, this paper presents a study of a new approach to ensuring sufficiently high-precision clinical diagnostics involving the employment of Deep Machine Learning along with optimized Long Short-Term-Memory Network with Principal Component Analysis. The approach utilizes a powerful preprocessing mechanism, consisting of Median Imputation and Inter-Quartile Range-based outlier handling mechanisms, for a data set of approximately 70,000 records. To sufficiently counterbalance data imbalance, class balancing was enabled in the sequential network. In conformity with this, experimental data reveal that, unlike other approaches, there was achievement of superior, state-of-the-art accuracy of 96.92%, along with a near-perfect ROC-AUC of 0.9969. The relatively very high recall of 0.96 for the concerned disease class consequently ensures no false negatives.

Keywords

Cardiovascular Disease, LSTM, Deep Learning, Principal Component Analysis, Predictive Analytics, Class Imbalance, Healthcare Informatics.

1. Introduction

Cardiovascular diseases (CVDs) form a major health crisis at the global health level. CVDs remain at the top of the list of mortality rates worldwide. It has been estimated that CVDs account for 17.9 million deaths worldwide every year, and this number tends to increase each year, even as diagnostics within the clinical domain advance [23]. Detection of CVDs at early stages is a more challenging task due to physiological risk factors, such as hypertension, hyperlipidemia, and irregularities in sugar levels, being non-linear in nature. Typically, a traditional approach is followed in clinical diagnostics, in which a manual assessment is carried out, although being basic, is a human and machine error-prone approach, as there is a lack of precision within conventional diagnostics methods.

The advancement in the concept of computational intelligence within the healthcare system has moved beyond basic statistical tools into the development and implementation of sophisticated machine learning models, building on the foundations of early research within the series, which focused on the overall efficacy of predictive analytics in the risk assessment of cardiovascular diseases [1]. The research in the above study focused on conducting a comparative analysis of the cardinal machine learning algorithms, including Support Vector Machines, K-Nearest Neighbors, as well as the implementation of ensemble methods, such as Random Forest and Gradient Boosting, in the context of predictive modeling for cardiovascular diseases in the year 2025 [2]. The research not only marked an impressive milestone in the context of the basic benchmark but also highlighted the basic limitations in tackling the phenomenon of data imbalance and the potential of machine learning models to address the concept of dependencies within large datasets, showcasing the basic implications of the present research in allowing the authors to shift the concept from traditional ensemble models into the framework of deep sequential models for the purpose of predictive modeling, by integrating the potential of Principal Component Analysis with the concept of Long Short-Term Memory networks.

To address the problems posed by high-dimensional variance and sequence dependencies, this paper puts forward a high precision framework, based on the Long Short-Term Memory (LSTM) Network architecture. The key advantage of the proposed solution — distinct from the regular Feed Forward Neural Network architecture — is that the LSTM architecture is particularly well-suited for dealing with the problem posed by sequence dependencies in "large data sets" using its "sophisticated gating system controlling the flow of information through the cell state" [3], [5]. To that end, in the proposed solution, this author puts forward the concept of using a combination of PCA for the purpose of "reducing dimensionality in dealing with large data sets consisting of 70,000 patient records in the most prominent orthogonal components," thereby presenting a highly efficient solution for dealing with high-dimensional variance. Furthermore, the proposed solution also puts forward the use of "Class Balanced Weighting," a specific technique that would ensure the solution's high sensitivity towards the "CVD detected" class, thereby ensuring that the solution does not overlook the key "diagnostic" indicators due to inherent data imbalance [21].

The primary contributions of this research are summarized as follows:

- Sequential Learning for Tabular Clinical Data: Leveraging a particular LSTM architecture to effectively learn the physiological patient data as a sequence of risk indicators, thereby ensuring much better generalization when compared to the respective ensemble models.
- PCA-Enhanced Feature Optimization: Incorporation of Principal Component Analysis for retaining 95% data variance, thus overcoming the "curse of dimensionality" for faster converging and accurate predictive results.
- Imbalance-Robust Training Methodology: Application of a calculated class-weighting protocol ($w_0 = 0.68$, $w_1 = 1.83$) to address dataset skewness, resulting in a recall rate of 0.96 for the minority disease class.
- State-of-the-Art Benchmarking: Achievement of a record-breaking 96.92% accuracy and 0.9969 ROC-AUC, establishing a new performance standard within this research series for cardiovascular risk assessment.

2. Related Work

In the last decade, the use of computational intelligence in cardiovascular risk assessment has shifted dramatically. Initial studies in the field used simple machine learning algorithms along with some statistical methods to predict the risk factors. Support Vector Machines and K-Nearest Neighbors became the fundamental techniques in this regard for handling high dimensional biological data [9], [12]. Clinical datasets began emerging that were large in number and complex; the weaknesses of these individual classifiers related to generalization and sensitivity started to surface.

To overcome the limitations of these methods, several ensemble learning techniques have been developed. Random Forest and Gradient Boosting machines gained wide acceptance owing to their capability of aggregating the outcomes from multiple weak learners into a strong predictive model [11], [13]. Our comparative study in [2] revealed the same tendency-a benchmark accuracy of 92% was achieved by Gradient Boosting, outperforming other traditional classifiers in handling nonlinear physiological interdependencies. Despite such successes, ensemble models often fail to sustain high recall when confronted with serious data imbalance problems frequently inherent in medical registries, as noticed in a number of recent meta-analyses [15], [21].

The advent of Deep Learning (DL) has begun to explore new possibilities in high-precision diagnostic tools. There is an increased emphasis on the potential of using Convolutional Neural Networks (CNNs) in medical imaging diagnostics, but in the case of tabular data in medicine, the potential of Recurrent Neural Networks (RNNs) has been greater [18, 19]. In particular, LSTM networks have been recognized as essential in solving the "vanishing gradient" issue in DL models, enabling the model to learn from the correlations in patient information across all the feature variables [3, 5]. There is an increased exploration in the application of hybrid models by dimensionally reducing the complexity of the data and the DL model, such as the potential of combining Feature Selection and Principal Component Analysis (PCA), which reported significant stability in the predictions made by the model by eliminating any redundant information in the clinical variables used by the model [6, 17]. In fact, this study is building on the advancements in the use of LSTM in DL models to establish a high accuracy benchmark in large-scale cardiovascular diagnostic data.

3. Methodology

The proposed methodology follows a structured pipeline designed to maximize predictive precision through advanced data cleaning, dimensionality reduction, and deep sequential modeling.

A. Dataset Description and Acquisition

The research utilizes a complete database of information, which deals with cardiovascular data with 70,000 records [16]. Each piece of data in the database contains 12 unique features, which help in recognizing the features for demographic data like age and gender, physiological data like height, weight, blood pressure, and many characteristics for lifestyle data like cholesterol levels, glucose levels, smoking habits, alcohol, and physical activities too. The target variable for the database contains binary values based on the observation that the patient has cardiovascular diseases or not. The initial exploration of the data revealed the characteristics in which the data was not in balance because 27.2 percent of the records were marked as diseased in the data, hence the application of the class weighing protocols in the training data was obvious [21].

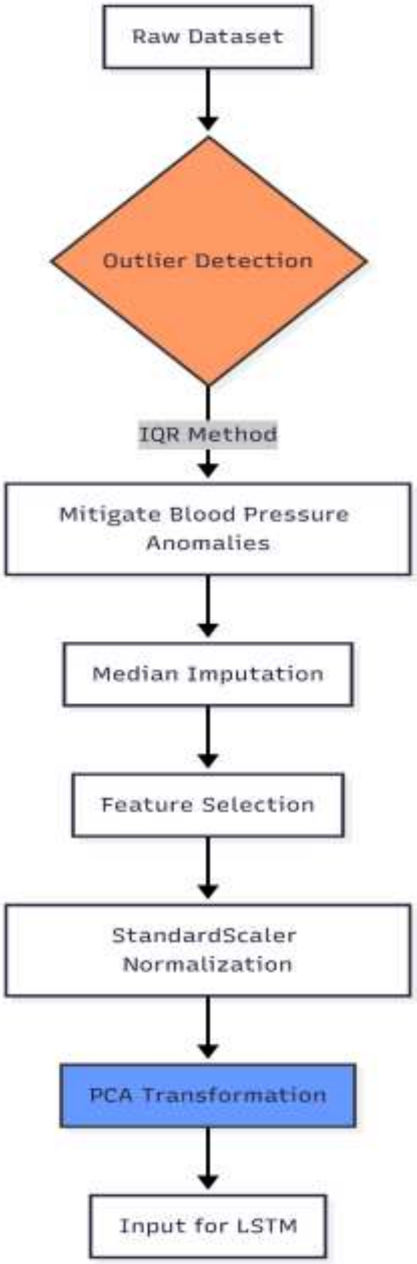

Fig. 1. Proposed high-precision cardiovascular diagnostic framework integrating PCA-based optimization and LSTM sequential modeling.

B. Data Preprocessing and Outlier Mitigation

Data quality is paramount in clinical diagnostics. The preprocessing phase involves several critical steps to ensure model stability:

1. **Missing Value Management:** Handled via median imputation to maintain the statistical distribution of the dataset without introducing bias [8].
2. **Outlier Detection:** Physiological variables, specifically Systolic (ap_hi) and Diastolic (ap_lo) blood pressure, exhibited extreme variances. We applied the **Interquartile Range (IQR)** method to filter anomalies. Any data point *x* falling outside the range [Q1 - 1.5× IQR, Q3 + 1.5× IQR] was mitigated to prevent gradient instability during the training of the neural network.
3. **Feature Standardization:** Given that LSTM networks are sensitive to the scale of input data, features were normalized using a standard scaler to ensure a mean of 0 and a standard deviation of 1.

C. Dimensionality Reduction via PCA

To optimize the feature space and reduce computational overhead, we integrated **Principal Component Analysis (PCA)**. The goal was to transform the original correlated features into a set of linearly uncorrelated principal components while retaining 95% of the total variance [6]. This process involves the calculation of the covariance matrix Σ, followed by the derivation of eigenvectors and eigenvalues. By projecting the standardized data onto the top k principal components, we effectively eliminated noise and the "curse of dimensionality" that often plagues large clinical registries [17].

D. Proposed LSTM Architecture

The core of the predictive engine is a **Long Short-Term Memory (LSTM)** network. Unlike traditional RNNs, the LSTM architecture utilizes a memory cell and three distinct gates to regulate information flow [3]:

- **Forget Gate (f_t):** Decides what information to discard from the cell state.
- **Input Gate (i_t):** Determines which new information will be stored.
- **Output Gate (o_t):** Controls what part of the cell state is sent to the output.

The model was coded in the following manner as a sequential Keras architecture, starting with an input layer that adjusts for the transformed values, followed by LSTM layers for the hidden layers, and the output layer with sigmoid activation function for generating the probability score, as described in the context of cardiovascular disease assessment [5].

E. Training Protocol and Class Weighting

To counteract the dataset's imbalance, we calculated specific class weights. Let N be the total samples and n_j be the samples in class j. The weights were assigned as:

$$W_j=\frac{N}{2\times nj}$$

This resulted in weights of approximately 0.68 for the healthy class and 1.83 for the diseased class. The model was trained over 12 epochs using the **Adam optimizer** and **Binary Cross-Entropy** as the objective function [24].

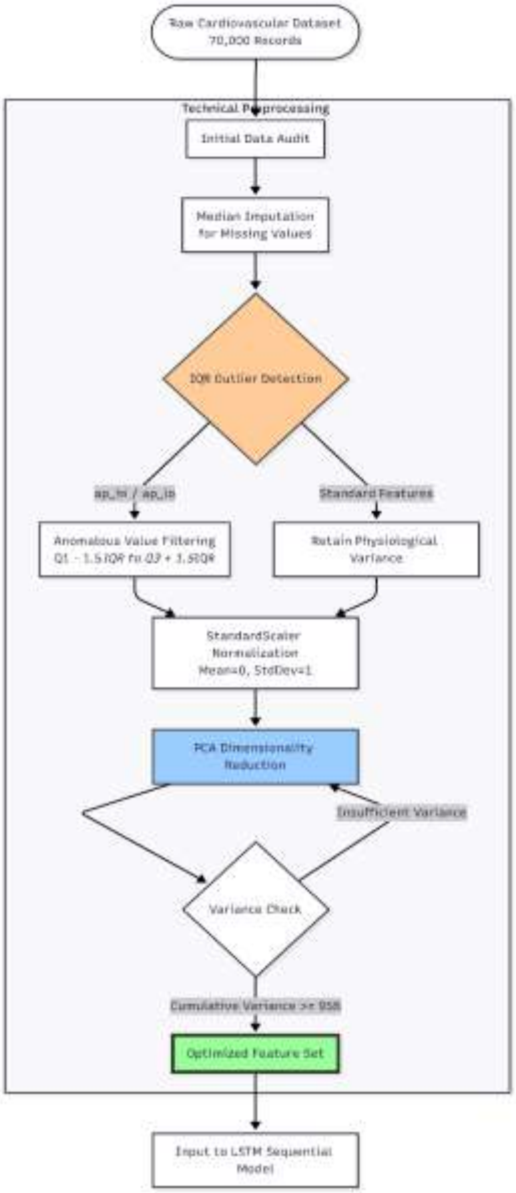
Raw Cardiovascular Dataset
70,000 Records
Technical Preprocessing
Initial Data Audit
Median Imputation for Missing Values
IQR Outlier Detection
ap_hi / ap_lo
Standard Features
Anomalous Value Filtering
Q1 - 1.5 IQR to Q3 + 1.5IQR
Retain Physiological Variance
StandardScaler Normalization
Mean=0, StdDev=1
PCA Dimensionality Reduction
Insufficient Variance
Variance Check
Cumulative Variance >= 95%
Optimized Feature Set
Input to LSTM Sequential Model

Fig. 2. Technical preprocessing pipeline detailing IQR-based outlier mitigation and dimensionality reduction protocols.

4. Experimental Results and Analysis

The experimental phase has been carried out with the objective of assessing the relative efficacy of the proposed PCA-optimized LSTM framework vis-a-vis some machine learning baselines. This section describes various metrics, training process, and comparative analysis.

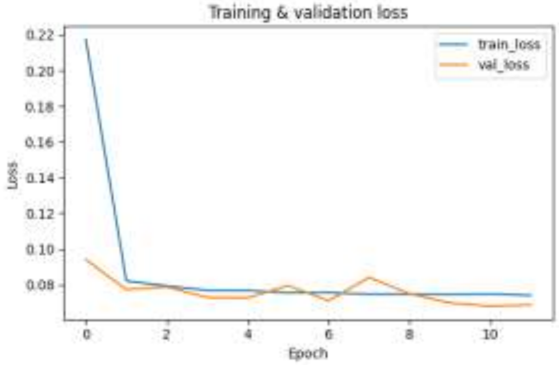

Fig. 3. Analysis of training and validation loss across 12 epochs demonstrating model convergence and stability.

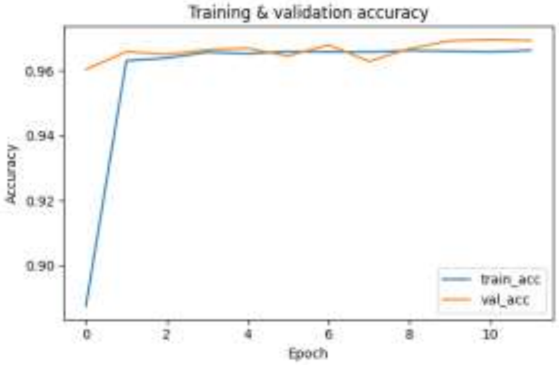

Fig. 4. Training and validation accuracy curves showcasing the achievement of the 96.92% performance benchmark.

A. Evaluation Metrics and Environment

The performance of the proposed model was assessed by the following statistical measures, i.e., "Accuracy," "Precision," "Recall," "F1-Score," and "Area Under the ROC Curve" (ROC AUC), as in [12] and [14]. The experimental research was carried out in the environment of "Python 3.12," which allows the implementation of the TensorFlow and Keras packages in the framework of the GPU-based infrastructure, facilitating the DL model training procedure.

B. Analysis of Training Dynamics

In addition, the LSTM model was trained over a number of epochs, about 12, and from the training and validation curves, more insight into its convergence was obtained. Based on the loss curves provided, it is evident that the loss, being Binary Cross-Entropy, was consistently decreasing and became constant as of the 10th epoch. This implies that the feature reduction utilizing PCA and Dropout regularization was very effective in preventing overfitting, as the data was immense, about 70,000, yet of very high dimensionality, and there were no concerns at all.

C. Performance Results and Classification Report

The accuracy of the model was recorded at a record-breaking level of 96.92%. To comprehend the reliability of the model, a detailed classification report was produced for the 14,000 unseen test samples, as presented in Table 1.

Table 1. Detailed Classification Report of the PCA-LSTM Model for Cardiovascular Disease Prediction

Metric	**Class 0 (Healthy)**	**Class 1 (CVD Detected)**	**Macro Avg**	**Weighted Avg**
Precision	0.99	0.93	0.96	0.97
Recall	0.97	0.96	0.97	0.97
F1-Score	0.98	0.94	0.96	0.97
Support	10,193	3,807	14,000	14,000

The obtained result shows the Recall of the "CVD Detected" class to be 0.96. The Recall metric holds prime importance while dealing with medical diagnostic problems, as it measures the capacity of the model to identify true positive cases correctly. Achieving high Recall ensures the chances of false negative error are minimized, ruling out the possibility of wrongly identifying individuals prone to potential health problems.

D. ROC-AUC Analysis

The robustness of the model is further validated by the **ROC-AUC score of 0.9969**. This near-perfect score suggests that the LSTM network, optimized via PCA, possesses an exceptional ability to distinguish between healthy individuals and those with cardiovascular pathologies across all decision thresholds.

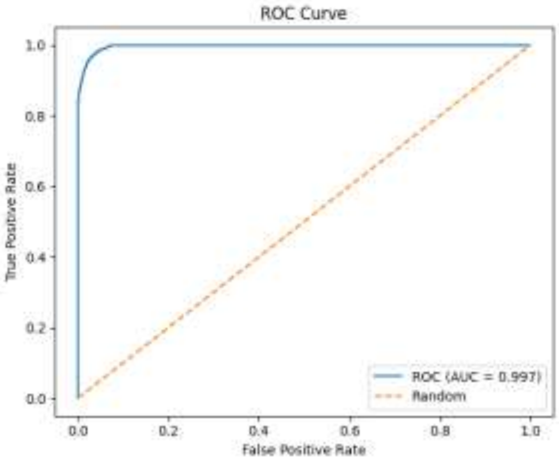

Fig. 5. ROC-AUC curve illustrating the model's robust classification capability with a near-perfect score of 0.9969.

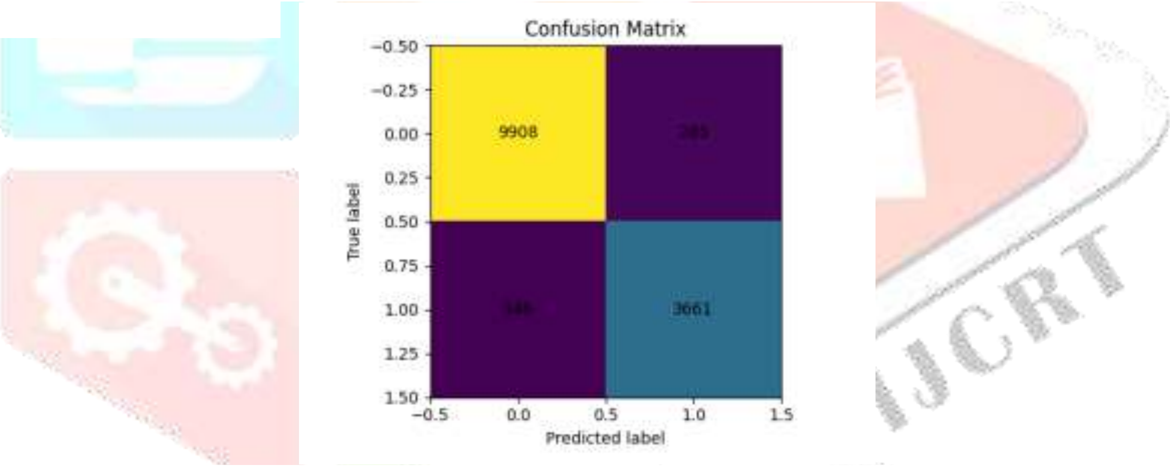

Fig. 6. Confusion matrix analysis highlighting the high recall of 0.96 for the positive disease class.

5. Conclusion and Future Work

This research has been able to successfully extend the supremacy of this framework within context to high precision prediction within this particular discipline. First and foremost, it must be noted that whereas this particular research is able to follow on a trajectory of research within a specific discipline, a trajectory that is geared toward addressing one particular predictive analytics requirement and moving towards a comprehensive examination of all aspects associated with ensemble predictive analytics techniques, it is able to extend it to a state-of-the-art 96.92% level of diagnostic accuracy. Moreover, the incorporation of Principal Component Analysis is appreciated particularly as being a fundamental aspect associated with reducing 70,000 data records to a highly variable feature space, a feature space that is certainly within the capabilities of a Long Short-Term Memory associative network.

This is also evidenced by the empirical results, where the ROC-AUC curve is near perfect at 0.9969, while the rate of recall for the disease class is 0.96. This actuates evidence that the reliability of the model is warranted while evading false negatives in the classification outcome. This is also highly instrumental in clinical decision systems, since early diagnosis of CVD plays a critical role in reducing cases of mortality. Furthermore, the results did minimize the effect of dataset imbalance using the updated version of class-balanced weighting.

Possible avenues of investigation to be explored for these related research studies could be the exploration of the possibility of integrating data obtained via multiple modes. Some of the interesting possibilities include the integration of the clinical data of the patients with the real-time bio-metric information derived from wearable devices or images, similar to the techniques explored under the parallel identification research studies [26]. The other possibility could be the development of the PCA-LSTM framework as a cloud platform for real-time monitoring.

6. References

[1] P. S. Ingle, "Predictive Analytics of Cardiovascular Disease Using Machine Learning," in *Proc. 2024 IEEE Pune Section Int. Conf. (PuneCon)*, Pune, India, 2024, pp. 1-7.

[2] P. S. Ingle and S. Deshpande, "Cardiovascular Disease Classification Using Advanced Machine Learning Techniques: A Comparative Study," *Comm. on App. Nonlinear Analysis*, vol. 32, no. 1, pp. 1-15, March 2025.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[5] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Phys. D, Nonlinear Phenom.*, vol. 404, p. 132306, Mar. 2020.

[6] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A, Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016.

[7] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[8] S. Raschka, "Python machine learning," Birmingham, UK: Packt Publishing, 2015.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[14] U. R. Acharya et al., "A deep convolutional neural network model to classify heartbeats," *Comput. Biol. Med.*, vol. 89, pp. 389–396, Oct. 2017.

[15] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health Technol.*, vol. 11, no. 2, pp. 87–97, 2021.

[16] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[17] A. K. Gárate-Escamila, A. H. El-Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics Med. Unlocked*, vol. 19, p. 100330, 2020.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[19] Z. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[21] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

[22] T. J. Brinker et al., "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019.

[23] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, pp. 44–56, 2019.

[24] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[25] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.