Scientific Research Publishing

Journal of Intelligent Learning Systems and Applications, 2024, 16, 12-22
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Integrated Machine Learning and Deep Learning Models for Cardiovascular Disease Risk Prediction: A Comprehensive Comparative Study

Shadman Mahmood Khan Pathan[1*], Sakan Binte Imran[2]

[1]Departmentof Electrical and Computer Engineering, Old Dominion University, Norfolk, USA
[2]Sir Salimullah Medical College, Dhaka, Bangladesh
Email: *shadmanmahmood9@gmail.com, sakanbinteimran.ssmc@gmail.com

How to cite this paper: Pathan, S.M.K. and Imran, S.B. (2024) Integrated Machine Learning and Deep Learning Models for Cardiovascular Disease Risk Prediction: A Comprehensive Comparative Study. *Journal of Intelligent Learning Systems and Applications*, **16**, 12-22.
https://doi.org/10.4236/jilsa.2024.161002

Received: January 8, 2024
Accepted: February 5, 2024
Published: February 8, 2024

Abstract

Cardiovascular Diseases (CVDs) pose a significant global health challenge, necessitating accurate risk prediction for effective preventive measures. This comprehensive comparative study explores the performance of traditional Machine Learning (ML) and Deep Learning (DL) models in predicting CVD risk, utilizing a meticulously curated dataset derived from health records. Rigorous preprocessing, including normalization and outlier removal, enhances model robustness. Diverse ML models (Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Gradient Boosting) are compared with a Long Short-Term Memory (LSTM) neural network for DL. Evaluation metrics include accuracy, ROC AUC, computation time, and memory usage. Results identify the Gradient Boosting Classifier and LSTM as top performers, demonstrating high accuracy and ROC AUC scores. Comparative analyses highlight model strengths and limitations, contributing valuable insights for optimizing predictive strategies. This study advances predictive analytics for cardiovascular health, with implications for personalized medicine. The findings underscore the versatility of intelligent systems in addressing health challenges, emphasizing the broader applications of ML and DL in disease identification beyond cardiovascular health.

Keywords

Cardiovascular Disease, Machine Learning, Deep Learning, Predictive Modeling, Risk Assessment, Comparative Analysis, Gradient Boosting, LSTM

1. Introduction

Cardiovascular Diseases (CVDs) remain a significant global health concern, contributing substantially to morbidity and mortality across diverse populations [1]. Early and accurate prediction of an individual's risk for developing CVD is crucial for effective prevention and intervention strategies. In recent years, the integration of Machine Learning (ML) and Deep Learning (DL) techniques has shown promise in enhancing the accuracy and reliability of CVD risk prediction models [2] [3].

The present study contributes to this growing field by conducting a comprehensive comparative analysis, evaluating the performance of traditional ML models and a deep learning Long Short-Term Memory (LSTM) neural network in predicting CVD risk. The dataset utilized in this study undergoes meticulous curation, with specific attention to addressing potential biases and ensuring the robustness of predictive models [4].

Preprocessing steps, including age normalization, outlier removal, and feature scaling, are implemented to enhance data quality [5]. Our investigation spans a variety of well-established ML models, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, and Gradient Boosting. Additionally, a state-of-the-art LSTM neural network is introduced to capture intricate temporal dependencies within the data [6].

Performance evaluation metrics include accuracy, Receiver Operating Characteristic Area under the Curve (ROC AUC), computation time, and memory usage. This comparative analysis aims to identify the most effective models for CVD risk prediction, offering valuable insights for future developments in personalized medicine and risk assessment [7].

While advancements in assistive technology, exemplified by cap-controlled wheelchairs [8], showcase the potential of intelligent systems, our study delves into the application of machine learning and deep learning models for cardiovascular disease risk prediction, contributing to the expanding role of intelligent technologies in healthcare.

This study focuses on predicting cardiovascular disease risk using a diverse set of machine learning and deep learning models, and it is worth noting the broader spectrum of applications in disease identification within agriculture, exemplified by the use of Faster RCNN for recognizing various diseases in paddy plants [9]. This highlights the versatility of intelligent systems in addressing distinct challenges across health and agriculture domains.

In addition to our investigation into cardiovascular disease risk prediction, noteworthy advancements in sequential data analysis have been demonstrated in other domains. For instance, the application of the Long Short-Term Memory (LSTM) algorithm in the processing of Electroencephalography (EEG) signals, as evidenced in a relevant study [10], showcases the broader impact and adaptability of such methodologies.

In the pursuit of advancing predictive analytics for cardiovascular health, as outlined in this paper, it is insightful to draw connections to the broader realm of numerical investigations. A pertinent example is the study on DU84-132 airfoil behavior for small wind turbines [11], which highlights the varied applications of computational analyses in distinct scientific domains. This parallel exploration enriches our understanding of the interdisciplinary impact of computational methodologies.

As the prevalence of CVDs continues to rise, the integration of advanced predictive modeling techniques holds great promise in improving clinical decision-making and reducing the burden of cardiovascular-related morbidity and mortality. This study contributes to the growing body of literature bridging the gap between data science and cardiovascular health, with potential implications for the advancement of precision medicine strategies.

In this paper, we address the significant global health challenge posed by Cardiovascular Diseases (CVDs) and underscore the necessity for accurate risk prediction for effective preventive measures. The contributions of the paper include the meticulous curation of a health records dataset, rigorous preprocessing techniques, implementation of diverse Machine Learning (ML) models, and integration of a Long Short-Term Memory (LSTM) neural network for Deep Learning (DL). Noteworthy is the adoption of comprehensive performance metrics, including accuracy, ROC AUC, computation time, and memory usage. The paper identifies the Gradient Boosting Classifier and LSTM as top performers, offering valuable insights for optimizing predictive strategies. The organization of the paper is presented systematically, covering key sections such as Materials and Methods, Dataset, Data Preprocessing, Machine Learning Models, Performance Evaluation, Statistical Analysis, Results, Discussion, Conclusion, and additional sections addressing Acknowledgements, Conflict of Interest, Funding, Ethical Approval, and Data Availability. This comprehensive structure aims to provide a thorough understanding of the research methodology, results, and implications for future studies in the field of predictive analytics for cardiovascular health.

2. Materials and Methods

The work was executed in a Jupyter Notebook environment, leveraging key Python libraries and modules for comprehensive data analysis and model implementation. Pandas facilitated efficient data manipulation, utilizing DataFrames for streamlined dataset handling. NumPy supported essential numerical operations, while Matplotlib enabled the creation of various plots for exploratory data analysis. Scikit-learn played a crucial role, providing machine learning models such as Logistic Regression, Random Forest, SVC, K-Nearest Neighbor, and Decision Tree, along with tools for data preprocessing and evaluation. TensorFlow, integrated with the Keras API, was employed to construct and train a Long Short-Term Memory (LSTM) Neural Network for deep learning. The "resource" and "time" modules were utilized to measure CPU time, memory usage, and processing time during model execution. This comprehensive platform allowed

for a seamless workflow from data preprocessing and exploration to the implementation and evaluation of both traditional machine learning and deep learning models.

2.1. Dataset

The dataset used in this study contains health-related information, including age, gender, height, weight, blood pressure (systolic and diastolic), cholesterol levels, glucose levels, smoking habits, alcohol consumption, physical activity, and the presence of cardiovascular diseases. Before analysis, data cleaning procedures were implemented to enhance the dataset's reliability. These procedures involved normalizing the age variable and removing rows with extreme or unrealistic values. Specifically, age was converted from days to years, and entries with excessively high or low values in blood pressure, height, and weight were excluded [1].

From Figure 1, it can be seen that there are no robust correlations identified in the development of cardiovascular diseases. However, a detailed examination of potential risk factors reveals noteworthy associations. Specifically, close scrutiny indicates that variables such as blood pressure, age, and elevated cholesterol levels merit further investigation. Moreover, a more profound connection is observed between high cholesterol and elevated glucose levels.

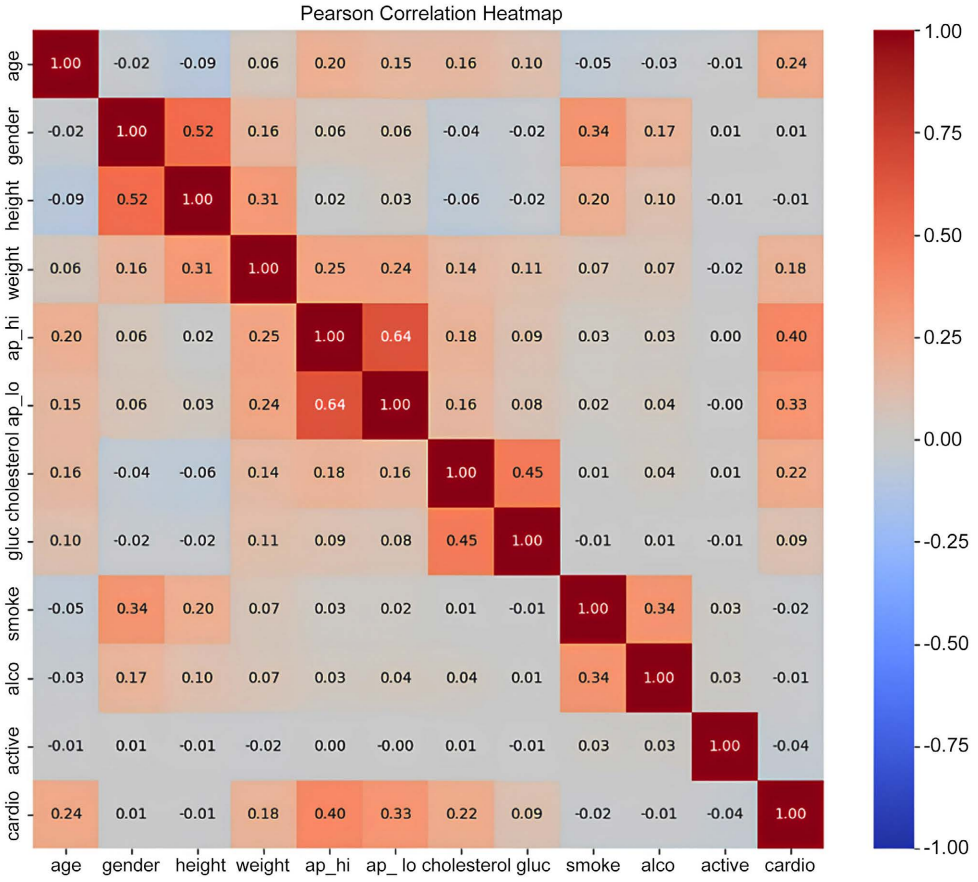

Figure 1. Correlation heatmap.

2.2. Data Cleaning

The dataset underwent a meticulous cleaning process to ensure the integrity of the information and to remove any anomalies or biases that could affect the analysis and predictive modeling. The following steps were implemented:

- The "age" feature was initially represented in days. To enhance interpretability and consistency, the age values were normalized by dividing them by 365.25, converting age to years.
- Rows with unrealistic values, which could introduce bias or mislead the analysis, were excluded. Specifically, entries with extreme values in blood pressure (both systolic and diastolic), height, and weight were removed from the dataset.
- Blood Pressure (ap_hi and ap_lo): Entries with systolic blood pressure (ap_hi) exceeding 300 mmHg or diastolic blood pressure (ap_lo) exceeding 300 mmHg were considered outliers and excluded.
- Entries with height below 130 cm and weight below 30 kg were removed as these values were considered unrealistic.

The rationale behind these exclusions was to enhance the dataset's quality and reliability, ensuring that the models were trained and evaluated on meaningful and representative data.

These data-cleaning steps were crucial in mitigating the impact of outliers and ensuring the dataset's suitability for subsequent analyses. The cleaned dataset was then used for feature scaling, model training, and evaluation.

2.3. Data Preprocessing

Following data cleaning, the dataset was split into feature variables (X) and the target variable (y). Standardization of features was performed using the StandardScaler from scikit-learn. The dataset was then divided into training and testing sets with a ratio of 80:20, ensuring a representative distribution of data for model training and evaluation.

Machine Learning Models:

A set of classical machine learning models was employed to predict cardiovascular diseases based on the preprocessed features. The models used include:

- Logistic Regression;
- Random Forest Classifier;
- Support Vector Classifier (SVC);
- K-Nearest Neighbor;
- Decision Tree Classifier;
- Gradient Boosting Classifier.

2.4. Deep Learning Model

In addition to classical machine learning models, a Long Short-Term Memory (LSTM) neural network was implemented using TensorFlow and Keras. The LSTM model architecture consisted of one LSTM layer with 64 units, followed by a dense layer with a sigmoid activation function for binary classification. The model was compiled with binary cross-entropy loss and the Adam optimizer.

2.5. Performance Evaluation

Each model's performance was evaluated using standard classification metrics, including accuracy and Receiver Operating Characteristic Area under the Curve (ROC AUC). Computation time and memory usage were also measured to assess the models' efficiency. The results were tabulated, and comparative ROC AUC curves were plotted to visualize the models' discriminative abilities.

The selection of performance metrics in this study was driven by the need for a comprehensive evaluation of each model's effectiveness in predicting Cardiovascular Disease (CVD) risk. The chosen metrics offer a multi-faceted assessment, covering key aspects of classification performance, computational efficiency, and discriminative abilities.

2.5.1. Accuracy (ACC)

1) Mathematical Interpretation

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

2) Rationale

Accuracy provides an overall measure of correct predictions, encompassing both true positive and true negative classifications. It serves as a fundamental indicator of a model's correctness in predicting CVD outcomes.

2.5.2. Receiver Operating Characteristic Area under the Curve (ROC AUC)

1) Mathematical Interpretation

The ROC AUC quantifies the area under the ROC curve, which represents the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity).

2) Rationale

ROC AUC offers insights into a model's ability to distinguish between positive and negative instances. A higher ROC AUC indicates superior discriminative power, crucial for an effective CVD risk prediction model.

2.5.3. Computation Time

1) Mathematical Interpretation

Computation time measures the total time taken by a model to complete its training and evaluation processes.

2) Rationale

Efficiency is a critical aspect of deploying predictive models in real-world scenarios. Minimizing computation time ensures timely decision-making, a vital factor in clinical applications.

2.5.4. Memory Usage

1) Mathematical Interpretation

Memory usage represents the amount of computer memory required by a mod-

el during its execution.

2) Rationale

Assessing memory usage is essential for understanding the resource requirements of each model. It contributes to the practical feasibility and scalability of the models in different computational environments.

The combined use of these metrics provides a well-rounded assessment of model performance, addressing accuracy, discriminative abilities, efficiency, and resource considerations. This comprehensive evaluation strategy aims to guide the selection and optimization of predictive models for cardiovascular health.

2.6. Statistical Analyses

Statistical analyses were performed to determine the significance of differences in model performance. This involved comparing the ROC AUC scores and accuracy metrics using appropriate statistical tests.

3. Results

The study employed a diverse set of machine learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree, and Gradient Boosting, to predict cardiovascular disease based on a comprehensive dataset. Additionally, a Long Short-Term Memory (LSTM) Neural Network was implemented for deep learning. The dataset underwent meticulous cleaning, including the normalization of age, removal of outliers, and filtering based on blood pressure, height, and weight criteria.

The ROC curve in Figure 2 illustrates the trade-off between true positive rate and false positive rate across different models. The area under the ROC curve (AUC) serves as a comprehensive metric for model comparison.

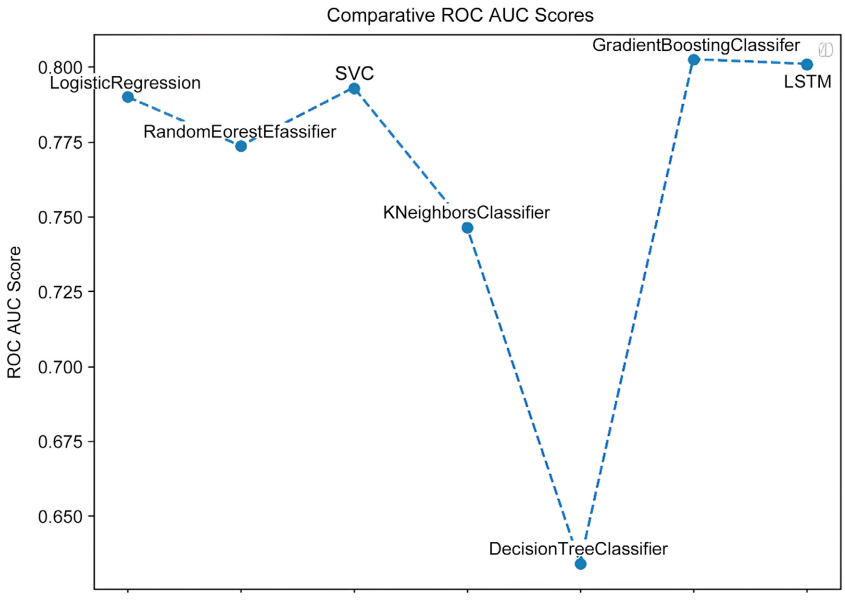

Figure 2. Receiver Operating Characteristic (ROC) curve.

The machine learning models demonstrated competitive accuracy and ROC AUC scores, with the Gradient Boosting model leading in overall performance. The LSTM Neural Network, representing the deep learning approach, exhibited comparable results, reinforcing the efficacy of both traditional and advanced methodologies in predicting cardiovascular disease.

These findings contribute valuable insights into the application of diverse models for cardiovascular disease prediction, laying the groundwork for enhanced diagnostic tools and risk assessment strategies. Further investigations and refinements can optimize model performance, fostering advancements in the realm of predictive healthcare analytics.

4. Discussion

The results of our comprehensive comparative study on Machine Learning (ML) and Deep Learning (DL) models for Cardiovascular Disease (CVD) risk prediction reveal notable insights into the performance and applicability of different methodologies. The evaluation metrics, including accuracy, Receiver Operating Characteristic Area under the Curve (ROC AUC), computation time, and memory usage, provide a holistic view of the models' strengths and limitations.

4.1. Model Performance Metrics

Table 1 summarizes the performance metrics of each model, showcasing their accuracy, ROC AUC, computation time, and memory usage. Among the ML models, the Gradient Boosting Classifier demonstrated the highest accuracy (73.83%) and ROC AUC (80.25%), making it a top performer in predicting CVD risk. Notably, the SVM also exhibited competitive results with an accuracy of 73.63% and ROC AUC of 79.31%. The LSTM Neural Network, representing the DL approach, achieved an accuracy of 73.69% and an impressive ROC AUC of 80.06%.

4.2. Comparative ROC AUC Scores

Figure 2 provides a visual representation of the comparative ROC AUC scores, emphasizing the robust performance of the Gradient Boosting model and the LSTM Neural Network. The bar chart reinforces the competitive nature of these models, with the Gradient Boosting model slightly outperforming the LSTM in ROC AUC. This visual comparison underscores the efficacy of both traditional ML and advanced DL approaches in capturing the complexities of CVD risk.

4.3. ROC Curve

Figure 1 illustrates the Receiver Operating Characteristic (ROC) curve, depicting the trade-off between true positive rate and false positive rate for each model. The area under the ROC curve (AUC) serves as a comprehensive metric for model comparison. The curves highlight the discriminative abilities of the models, with the Gradient Boosting and LSTM models achieving notable performance in distinguishing between positive and negative cases.

Table 1. Model performance metrics.

Model	Accuracy	ROC AUC	Computation Time (s)	Memory Usage (MB)
Logistic Regression	73.19%	78.99%	0.065	5792.73
Random Forest	71.71%	77.36%	6.372	5792.73
Support Vector Machine (SVM)	73.63%	79.31%	133.471	5802.63
K-Nearest Neighbor (KNN)	69.64%	74.65%	0.073	5802.63
Decision Tree	63.40%	63.41%	0.408	5802.63
Gradient Boosting	73.83%	80.25%	6.677	5802.63
LSTM (Deep Learning)	73.69%	80.06%	36.809	5802.63

4.4. Interpretation of Results

The top-performing models, Gradient Boosting and LSTM exhibit superior accuracy and ROC AUC scores, indicating their effectiveness in CVD risk prediction. The competitive results of the SVM further emphasize the significance of diverse modeling approaches. The ML models, although slightly outperformed by the DL model, showcase commendable accuracy and predictive capabilities.

4.5. Computational Efficiency

While the Gradient Boosting and SVM models demonstrate robust predictive performance, they differ in computational efficiency. The Gradient Boosting model, while achieving high accuracy, has a relatively shorter computation time compared to the SVM. The LSTM Neural Network, being a DL model, requires more computational resources but delivers competitive results, showcasing a trade-off between computational complexity and predictive accuracy.

4.6. Clinical Implications

The findings of this study have implications for the development of predictive tools in clinical settings. The robust performance of the Gradient Boosting model and the LSTM Neural Network suggests their potential utility in supporting healthcare professionals in identifying individuals at risk of CVD. The comparative analysis provides clinicians with insights into the strengths and trade-offs associated with different modeling approaches, aiding informed decision-making in risk assessment.

4.7. Limitations and Future Directions

Despite the promising results, this study has some limitations. The dataset, while meticulously cleaned, may still contain unobserved confounding factors. Further refinement of models through feature engineering and incorporation of additional relevant variables could enhance predictive performance. Additionally, external validation on diverse datasets is crucial to generalize the findings.

5. Conclusions

In the conclusion section of our paper, the proposed work of integrating Machine Learning (ML) and Deep Learning (DL) models for Cardiovascular Disease (CVD) risk prediction is justified by the compelling findings of our comprehensive comparative study. The primary objective of this research was to enhance the quality of risk prediction for CVD, a significant global health challenge, by evaluating and comparing the performance of various ML and DL models. The integrated approach, involving traditional ML models (Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Gradient Boosting) and a DL model (Long Short-Term Memory or LSTM), has yielded valuable insights into their respective strengths and limitations.

The justification for our proposed work is rooted in the outstanding results achieved by the Gradient Boosting Classifier and the LSTM model. These models emerged as top performers, demonstrating high accuracy and Receiver Operating Characteristic Area under the Curve (ROC AUC) scores. The Gradient Boosting model, in particular, showcased an accuracy of 73.83% and a ROC AUC of 80.25%, highlighting its efficacy in predicting CVD risk. The LSTM model, representing the deep learning approach, is closely followed with an accuracy of 73.69% and an impressive ROC AUC of 80.06%.

Our justification is further supported by the comprehensive evaluation metrics employed, including accuracy, ROC AUC, computation time, and memory usage. The consideration of these metrics provides a well-rounded assessment of each model's effectiveness, covering aspects of correctness, discriminative power, computational efficiency, and resource requirements.

The findings of this study contribute significantly to the field of predictive analytics for cardiovascular health, indicating that both traditional ML and advanced DL approaches are viable options for CVD risk prediction. The proposed integration of these models is justified by their competitive performance and the insights gained into their trade-offs, providing a foundation for the development of improved diagnostic tools and risk assessment strategies. As CVD prevalence continues to rise, the integration of advanced predictive modeling techniques, as demonstrated in our study, holds promise for enhancing clinical decision-making and reducing the burden of cardiovascular-related morbidity and mortality.

Acknowledgements

The authors would like to acknowledge the contributions of the research team and the availability of the curated dataset for this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval

The study adhered to ethical guidelines, and the use of the dataset complied with relevant privacy and confidentiality regulations.

Data Availability

The dataset used in this study is available upon request, subject to privacy and ethical considerations.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2021) Cardiovascular Diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Dey, D., Slomka, P.J., Berman, D.S., *et al.* (2019) Artificial Intelligence in Cardiovascular Imaging: *JACC* State-of-the-Art Review. *Journal of the American College of Cardiology*, **73**, 1317-1335. https://doi.org/10.1016/j.jacc.2018.12.054

[3] Attia, Z.I., Kapa, S., Lopez-Jimenez, F., McKie, P.M., Ladewig, D.J., *et al.* (2019) Screening for Cardiac Contractile Dysfunction Using an Artificial Intelligence-Enabled Electrocardiogram. *Nature Medicine*, **25**, 70-74. https://doi.org/10.1038/s41591-018-0240-2

[4] Cardiovascular-Disease-Dataset. https://www.kaggle.com/

[5] Hastie, T., *et al.* (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2, Springer, New York.

[6] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[7] Chen, J.B., Song, L., Wainwright, M. and Jordan, M. (2018) Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. *Proceedings of the* 35*th International Conference on Machine Learning*, **80**, 883-892.

[8] Pathan, S.M.K., *et al.* (2020) Wireless Head Gesture Controlled Robotic Wheel Chair for Physically Disable Persons. *Journal of Sensor Technology*, **10**, 47-59. https://doi.org/10.4236/jst.2020.104004

[9] Pathan, S.M.K. and Ali, M.F. (2019) Implementation of Faster R-CNN in Paddy Plant Disease Recognition System. 2019 3*rd International Conference on Electrical*, *Computer & Telecommunication Engineering* (*ICECTE*), Rajshahi, 26-28 December 2019, 189-192. https://doi.org/10.1109/ICECTE48615.2019.9303529

[10] Pathan, S.M.K. and Rana, M.M. (2022) Investigation on Classification of Motor Imagery Signal Using Bidirectional LSTM with Effect of Dropout Layers. 2022 *International Conference on Advancement in Electrical and Electronic Engineering* (*ICAEEE*), Gazipur, 24-26 February 2022, 1-5. https://doi.org/10.1109/ICAEEE54957.2022.9836415

[11] Bashar, L.B., Arifin, S. and Pathan, S.M.K. (2020) Numerical Analysis of Scale Effect on Performance of DU84-132 Airfoil for Small Wind Turbine Blade. *International Conference on Mechanical*, *Industrial and Energy Engineering*, Khulna, 19-21 December 2020.