Vol. 9 No.1 Tahun 2023, pp 28 - 36
Jurnal Teknologi dan Manajemen Informatika
http:// http://jurnal.unmer.ac.id/index.php/jtmi
P-ISSN: 1693-6604 E-ISSN: 2580-8044

Enhancement of Coronary Heart Disease Prediction using Stacked Long Short Term Memory

Cinthiya[1], Raymond Sunardi Oetama[2]

Universitas Multimedia Nusantara, Indonesia

Info Artikel

Riwayat Artikel
Diterima: 27-02-2023
Direvisi: 29-05-2023
Disetujui: 07-06-2023

Kata Kunci
Penyakit Jantung Koroner,
Stacked LSTM,
Single LSTM,
Prediksi

✉ Corresponding Author
Raymond Sunardi Oetama,
Universitas Multimedia Nusantara,
Tel. +622154220808
Email raymond@umn.ac.id

ABSTRAK

Angka kematian yang tinggi akibat penyakit jantung koroner telah menjadi perhatian global dalam dunia kesehatan, di mana pasien dengan penyakit jantung koroner tidak lagi hanya orang dewasa dan lanjut usia, namun saat ini juga terjadi peningkatan jumlah kasus penyakit jantung koroner pada pasien yang berusia muda. Dalam membandingkan algoritma pembelajaran mesin pada dataset penyakit jantung *UCI*, akurasi tetap berada di bawah 90%, terutama menggunakan *Long short-term memory* (LSTM). Namun, sebuah studi lain berhasil mencapai akurasi di atas 90% dengan menggunakan algoritma *Stacked* LSTM dengan basis data yang berbeda. Untuk mencapai hal ini, tujuan penelitian ini adalah untuk melampaui performa *Single* LSTM serta meningkatkan akurasi prediksi penyakit jantung di atas 90% menggunakan algoritma *Stacked* LSTM pada dataset penyakit jantung *UCI*. Dengan menggunakan tiga lapisan dan teknik *dropout*, penelitian ini berhasil mencapai tingkat akurasi tertinggi sebesar 91,8% pada *epoch* ke-114, dengan tingkat kegagalan (loss) sebesar 20%. Melampaui performa *Single* LSTM, penelitian ini berhasil mencapai akurasi melebihi 90% yaitu 91,8%.

INTRODUCTION

Health and fitness are one of the keys to living a longer, more active, and pleasurable life, in other words, health is a wealth that each individual can and must preserve [1]. As a result, it is critical to continually keep a healthy body and seek early treatment if any ailment is suspected. Sustaining health is simple, for example, having a good and regular diet, frequent basic physical exercise such as walking, getting up from a chair, obtaining adequate hours of sleep, limiting alcohol intake, and avoiding smoking [2]. Many positive effects will be felt by the body both physically and mentally by managing the condition of the body to stay healthy and maintain ideal body weight, such as lowering the risk of heart disease, obesity/overweight, osteoporosis, sleep apnea, diabetes mellitus, asthma, high blood pressure and cholesterol, skin disorders, stress, and depression [3]. Cardiovascular diseases (CVD) are disorders that can cause difficulties with the heart and blood vessels and are mainly caused by unhealthy and fatty diets, cigarette smoking, obesity, and alcohol intake [4]. The heart is one of the most vital organs in the human body [5]. This is because the heart pumps blood through the blood arteries, carrying oxygen and other nutrients to other vital organs in the body. As a result, heart health impacts the health of other organs in the body [6].

Despite the importance of the heart organ in the human body, the World Health Organization estimates that in 2021, there will be approximately 8 million cases of death from CVD each year, accounting for approximately 31% of all causes of death worldwide, with this

figure expected to rise to 23 million by 2030 [7]. CVD encompasses a wide range of disorders, including coronary heart disease, heart attack, and heart failure [8]. Coronary heart disease is one of the most prevalent CVD-related illnesses in Southeast Asia, particularly in Indonesia, with a total mortality rate of 1.25 million individuals per 250 million inhabitants each year [9]. This puts coronary heart disease the leading cause of death in Indonesia, surpassing TB and stroke. Coronary heart disease is typically induced by a narrowing of the coronary arteries as a result of atherosclerosis (fat buildup on the coronary artery walls) [10]. Other external variables that promote coronary heart disease include elevated cholesterol levels, diabetes, obesity, lack of exercise, high blood pressure, age, and heredity [11]. Children are now at a higher risk of developing ischemic heart disease than adults [12].

The impact of Coronary heart disease will be felt by the sufferer when the blood flow that is supposed to carry oxygen to the heart muscle is blocked or reduced, causing the sufferer to feel chest pain and difficulty breathing. At a certain level, coronary heart disease can cause the sufferer to experience angina pectoris, a heart attack, and even death [8]. Given the high number of victims of coronary heart disease, it is critical to understand a person's chance of getting coronary heart disease [13]. This is critical so that in the future, coronary heart disease can be predicted and treated earlier with quicker and more accurate procedures [14].

A study was conducted to assess the prediction accuracy of various machine learning algorithms using the UCI heart disease dataset. The results, evaluated using a confusion matrix, revealed the following accuracy rates: Logistic Regression achieved 86.81%, SVM achieved 87.91%, KNN achieved 86.81%, Random Forest achieved 82.42%, Decision Tree achieved 78.02%, and Naive Bayes achieved 82.42% [15]. Although these accuracies are commendable, they fall short of 90%. In another study using the same dataset, different machine learning algorithms and ensemble models were evaluated for accuracy on a specific task. The results indicated that Random Forest achieved 83.6%, Support Vector Machine achieved 81.3%, K-Nearest Neighbors achieved 82.8%, GRU achieved 81.46%, and LSTM achieved 81.31% [16]. From this case, the accuracy of a single LSTM model remained below 90%. Similarly, a separate study found that the prediction accuracy of LSTM on the same dataset was only 84.56% [17]. Contrastingly, a notable study [18] reported accuracy rates exceeding 90% by transitioning from a Single LSTM algorithm to Stacked LSTM algorithms, leveraging different databases. Motivated by this improvement, the objective of our study is to enhance Heart Disease Prediction accuracy by adopting Stacked LSTM algorithms on the UCI heart disease dataset, surpassing the performance of Single LSTM algorithms to achieve an accuracy of more than 90%.

As shown in Figure 1, the Long Short Term Memory algorithm is a subset of the RNN (Recurrent Neural Network) that is very capable of long-term learning [19]. LSTM enables the network to retain what it learns over an extended amount of time. Due to the output gate creating a network loop, the LSTM design is merely recurrent. All gates, including the input gate, receive the output of the present timestamp and use it to calculate the following timestamp. It implies that the prior network state is not wasted and is used for the subsequent prediction [20].

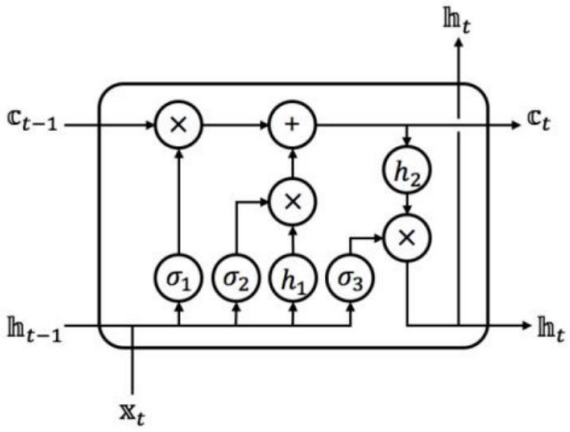
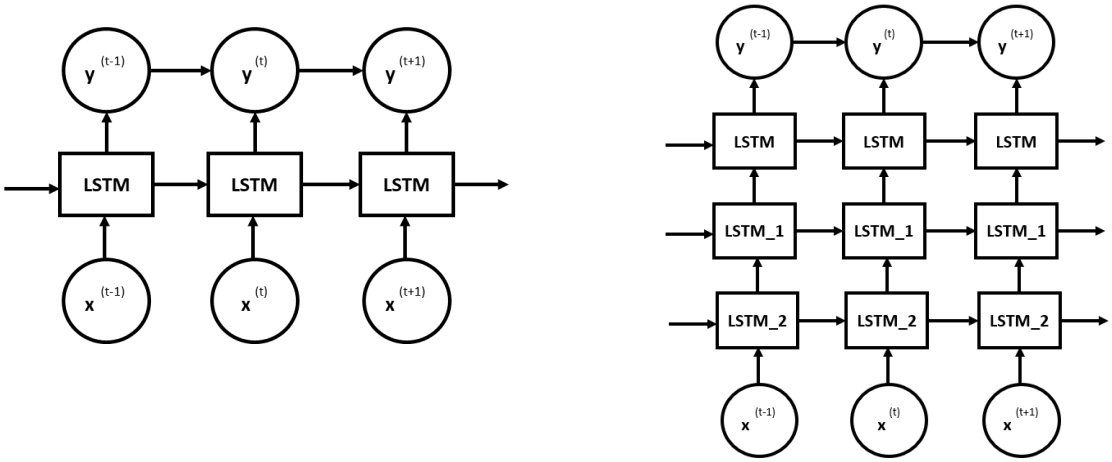


Figure 1. LSTM [20]

Unlike the Single LSTM Algorithm, Stacked LSTM is an algorithm that applies the LSTM architecture in a Stacked manner commonly defined as LSTM with a model consisting of several layers of LSTM, as shown in Figure 2. Stacked LSTM was developed and first used for voice detection. The Stacked LSTM model employs several levels of LSTM models that are Stacked before being transmitted to the end output. In Layered LSTM, the first LSTM model layer generates a sequence vector that is used as the input to the next LSTM model layer. Furthermore, Stacked LSTM gets input from the prior LSTM model, allowing the model to identify trends in the dataset used. It is also helpful for avoiding the chance of overfitting the dataset [21].



Single LSTM Stacked LSTM

Figure 2. Single LSTM Versus Stacked LSTM [20]

METHOD

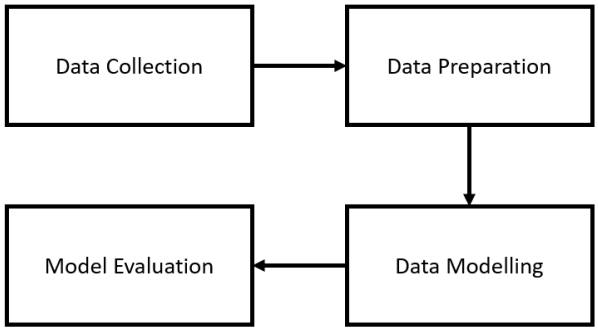


Figure 3. Research Framework

Figure 3 shows the Research Framework. This study only discusses the evaluation stage and will not proceed to the deployment stage.

1. The first stage is to collect data. This research employs a dataset named the heart disease dataset, which was obtained straight from the UCI Machine Learning Respiratory and is also known as the Cleveland heart disease dataset UCI [22]. This information was acquired by gathering patient data in a Cleveland Clinic Foundation database. Table 1 shows the dataset variables, and Table 2 shows the Descriptive Statistics for each variable. Descriptive data are provided for all factors, including the number of rows, mean, standard deviation (std), minimum, and maximum. All factors have 303 entries of data. Begun with an average age of 54, a standard deviation of 9, a lowest age of 29, and a highest age of 77. This chart is ended up with a variable Target, which has a mean of one, its standard variation and minimum are zero, and its maximum is one.

Table 1. Description and Data Types of Variables

Variable.	Description	Data Types
Age	Age in years	Integer
Sex	Gender (1 = male; 0 = female)	Binary
Cp	Type of chest pain (1: typical angina; 2: atypical angina ; 3: non-anginal pain; 4: asymptomatic)	Nominal
Trestbps	Resting blood pressure	Float
Chol	Cholesterol levels in mg/dl	Float
Fbs	Sugar level > 120 mg/dl (1 = true; 0 = false)	Binary
Restecg	Electrocardiographic results (0: normal; 1: abnormal; 2: possible ventricular hypertrophy)	Nominal
Thalach	Maximum heart rate record	Float
Exang	exercise-induced angina (1 = yes; 0 = no)	Binary
Oldpeak	ST depression	Float
Slope	ST segment (1: upsloping; 2: flat; 3: downsloping	Nominal
Ca	Major vessels (0-3) (marked with fluoroscopy)	Nominal
Thal	3 = normal; 6 = fixed defects; 7 = reversible defects	Nominal
Target	1 = yes, 0 = no	Binary

Table 2. Descriptive Statistics of Variables

Variable.	n	mean	std	minimum	maximum
Age	303	54	9	29	77
Sex	303	1	0	0	1
Cp	303	1	1	0	3
Trestbps	303	132	18	94	200
Chol	303	246	52	126	564
Fbs	303	0	0	0	1
Restecg	303	1	1	0	2
Thalach	303	150	23	71	202
Exang	303	0	0	0	1
Oldpeak	303	1	1	0	6
Slope	303	1	1	0	2
Ca	303	1	1	0	4
Thal	303	2	1	0	3
Target	303	1	0	0	1

2. The next stage is Data Preparation which will change the raw data that is not ready for use into information that is more useful and can be processed further. From data preparation to data modeling, all processes are performed using Python. In carrying

out the data preparation stage in this study, it consists of several other steps, as followings:

a. Data Cleansing is used to maintain the quality of the data used in the analysis, data cleansing is needed to find and eliminate variables and values that are either unused, irrelevant, or errors (missing values) [23]. The dataset used does not contain null (empty) values , it is clean and ready to be used.
b. Meanwhile, data splitting is done to divide the dataset into two parts consisting of training data and test data with a data range that is tailored to the needs and objectives of the research. In addition, at this stage also, the data train will be divided back into three parts, namely data training, data testing, and data validation. The dataset is divided into 80% data training and 20% data testing and afterward, data training is divided into two parts, namely 80% data training and 20% data validation.

3. The Modeling step will involve the development of a Stacked LSTM sequential model from TensorFlow using Python. Some of the parameters used are activation = 'sigmoid', loss = 'binary_crossentropy', optimizer = 'adam'.
4. The evaluation used is the measurements of the prediction models that have been built in the modeling stage obtained from the results of using the confusion matrix. These measurements include accuracy, F1-score, precision, and recall.

RESULT AND DISCUSSION

The structure of the Stacked LSTM model can be seen in Figure 4. There are three stacks which are lstm, lstm_1, and lstm_2, and three dropouts. Dropouts are applied to minimize overfitting [24]. As shown in Figure 5, the overfitting is reduced as the difference between accuracy and the accuracy validation becomes smaller. The total parameters are 49,201 and all of these parameters are trainable.

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 1, 48)	11904
dropout (Dropout)	(None, 1, 48)	0
lstm_1 (LSTM)	(None, 1, 48)	18624
dropout_1 (Dropout)	(None, 1, 48)	0
lstm_2 (LSTM)	(None, 48)	18624
dropout_2 (Dropout)	(None, 48)	0
dense (Dense)	(None, 1)	49

Total params: 49,201
Trainable params: 49,201
Non-trainable params: 0

Figure 4. Research Framework

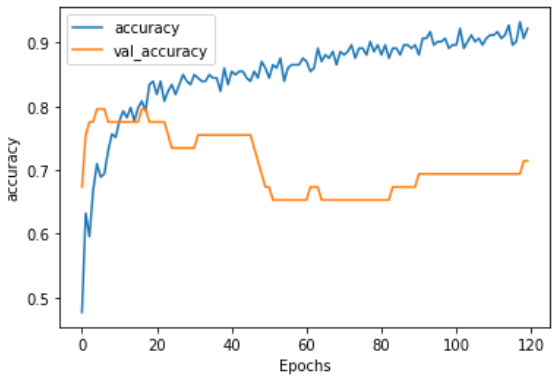


Figure 5. Accuracy versus Accuracy Validation on Epochs

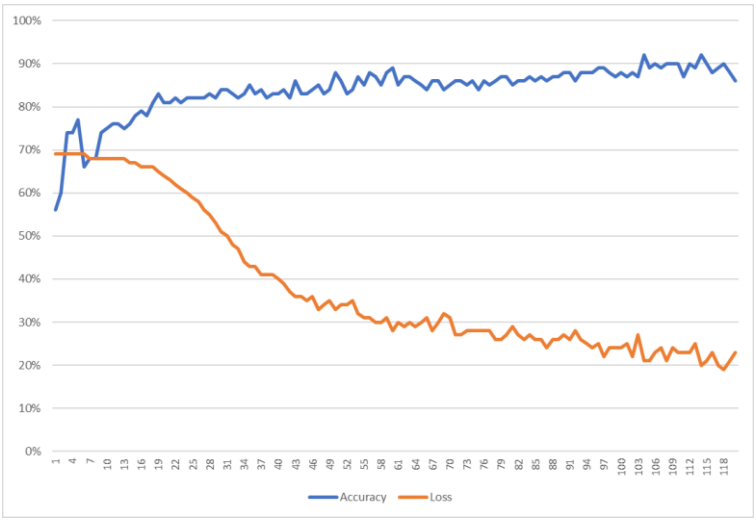


Figure 6. Accuracy and Loss Chart

Furthermore, as shown in Figure 6, the highest accuracy results from the 120 epochs that were run were at epoch 114 which was 92% with a loss was 20%. Furthermore, after the Stacked LSTM solution model using the data train has been successfully built, predictions will be made on test data from the dataset used.

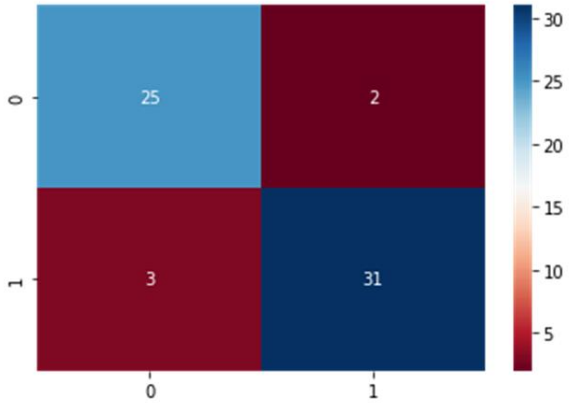


Figure 7. Confusion Matrix

As shown in Figure 7, the Confusion Matrix which applies the Stacked LSTM algorithm produces an accuracy of 91.80%, an F1 – Score of 92.53%, a Precision Score of 93.94%, and a Recall Score which is also 91.18%. Just like Single Long Short-Term Memory, the solution model built using the Stacked LSTM algorithm, can also be said to be able to predict coronary heart disease in general. This can be seen from the target data (actual) which is relatively the same as the predicted data generated by the model. There are as many as 25 "0" data that are predicted as "0" (True Positive), then as many as two "1" data which are predicted as "0" (False Positive), and as many as 3 "0" data are predicted as "1" (False Negative), as well as 31 data "1" which is predicted as "1" (True Negative).

Table 3. Comparison of The Study Result with Previous Research on the UCI Dataset

Articles	Data	Algorithms	Accuracy
"Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method" [16]	UCI Heart Disease, 303 rows, 14 features	Single LSTM	81,31%
Heart disease prediction based on random forest and LSTM [17]	UCI Heart Disease, 303 rows, 14 features	Single-LSTM	84.56%

This Study	UCI Heart Disease, 303 rows, 14 features	Stacked LSTM	91.8%

Compared with other studies in the same dataset as shown in Table 3, the first study titled "Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method" [16] focuses on heart disease prediction using a dataset from UCI with 303 rows and 14 features. The study utilizes a Single LSTM model and reports an accuracy of 81.31%. The second study, "Heart disease prediction based on random forest and LSTM" [17], also employs the UCI Heart Disease dataset with the same characteristics. The Single LSTM model achieves an accuracy of 84.56%. In this context, our study expands on the previous research by introducing a Stacked LSTM model for heart disease prediction. Using the same UCI Heart Disease dataset, the Stacked LSTM model achieves an impressive accuracy of 91.8%.

Furthermore, table 4 shows some ways to improve the prediction accuracy of LSTM from some previous studies. A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs" [18] focuses on atrial fibrillation prediction using ECG data. The study utilizes a Single LSTM model and reports an accuracy of 87%. However, with the implementation of a Stacked LSTM model, the accuracy is further improved to 92%. The potential for further enhancement in heart disease prediction by exploring alternative databases, such as ECG data. The utilization of ECG data is particularly advantageous as it involves time series information [25], and LSTM models have demonstrated efficacy in analyzing and extracting patterns from such data [26-28]. The last study [29] reports an accuracy of 90.13% for the Single LSTM model. Additionally, the implementation of a BiLSTM model leads to a slightly improved accuracy of 90.52%. It shows there is another opportunity to make improvements using another form of LSTM such as BiLSTM. Overall, in predicting heart diseases it suggested applying some improvement methods rather than just a single LSTM algorithm.

Table 4. Advancing Prediction Accuracy of Single LSTM

Articles	Data	Algorithms	Accuracy
“A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs” [18]	ECG data, 160 rows, time series data	Single LSTM Stacked LSTM	87% 92%
Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases [29]	12-lead ECG database	Single LSTM BiLSTM	90.13 90.52

CONCLUSION

By applying the Stacked LSTM algorithm, this study successfully enhanced the accuracy of heart disease prediction compared to a previous study that used the same dataset. The Stacked LSTM model consisted of three stacks and incorporated dropout techniques to mitigate overfitting. The highest accuracy achieved was 91.8% at epoch 114, with a corresponding loss of 20%. As a result, our study successfully achieves its objective of enhancing Heart Disease Prediction accuracy beyond 90% by implementing Stacked LSTM algorithms on the UCI heart disease dataset, surpassing the performance of Single LSTM algorithms. Following the completion of this study, it is highly recommended to investigate the use of another version of the LSTM algorithm for prediction, particularly in the field of health. Implement the enhanced model rather than relying solely on the single LSTM to achieve superior performance.

REFERENCES

[1] Ali, M. A., Kamraju, M., & Vani, M. (2017). Importance of health and fitness in life. *Int Fed Comp Sci Sports*, 41-45.

[2] New Zealand Government. (2018).A Guide for Maintaining Health and Wellbeing. Health and safety. Govt. Nz. Retrieved from https://www.healthandsafety.govt.nz/assets/Documents/A_Guide_for_Maintaining_Health_and_Wellbeing.

[3] Kumar, K. (2017). Importance of Healthy Life Style in Healthy Living. *Juniper Online J. Public Heal*, 2(5), from doi: 10.19080/jojph.2017.02.555596.

[4] World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(vcds)

[5] Wibisono, A. B. & Fahrurozi, A. (2019). Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner. *J. Ilm. Teknol. dan Rekayasa*, 24(3), 161–170, from doi: 10.35760/tr.2019.v24i3.2393.

[6] Rusdiana, T., Putriana, N. A., Sopyan, I. Gozali, D. and Husni, P. (2019). Pemberian Pemahaman Mengenai Sediaan herbal yang Berfungsi untuk Pemeliharaan Kesehatan jantung dan Ginjal di Desa Cibeusi, Sumedang, Jawa Barat. *J. Pengabdi. Kpd. Masy*, 4(6), 139–141.

[7] Kompaspedia. (2021) Hari Jantung Sedunia: Pentingnya Pencegahan Penyakit Jantung. Retrieved from https://kompaspedia.kompas.id/baca/paparan-topik/hari-jantung-sedunia-pentingnya-pencegahan-penyakit-jantung.

[8] NHS Service. (2021). Cardiovascular disease – NHS. Retrieved from https://www.nhs.uk/conditions/cardiovascular-disease.

[9] Oscar Primadi. (2021). Penyakit Jantung Penyebab Kematian Tertinggi, Kemenkes Ingatkan CERDIK – Sehat Negerik. Retrieved from https://sehatnegeriku.kemkes.go.id/baca/umum/20170801/2521890/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-2.

[10] Themistocleous, I., Stefanakis, M. & Douda, H. T. (2017). Coronary Heart Disease Part I : Pathophysiology and Risk Factors. *J. Phys. Act. Nutr. Rehabil.*, April, 167–175.

[11] Hajar, R. (2017). Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views*, 18(3), 109-114, from doi: 10.4103/Heartviews.Heartviews_106_17.

[12] De Ferranti, S. D., Steinberger, J., Ameduri, R., Baker, A., Gooding, H., Kelly, A. S., & Zaidi, A. N. (2019). Cardiovascular risk reduction in high-risk pediatric patients: a scientific statement from the American Heart Association. *Circulation*, 139(13), 603-634.

[13] Healthy people. (2021). *Heart Disease and Stroke Healthy People*. Retrieved from https://www.healthypeople.gov/2020/topics-objectives/topic/heart-disease-and-stroke.

[14] Riani, A., Susianto, Y., & Rahman, N. (2019). Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes. *Journal of Innovation Information Technology and Application (JINITA)*, 1(01), 25-34.

[15] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. Malaysian *Journal of Computer Science*, 132-148.

[16] Javid, I., Alsaedi, A. K. Z., and Ghazali, R. (2020). "Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method," *Int. J. Adv. Comput. Sci. Appl*., 11(3), pp. 540–551, from doi: 10.14569/ijacsa.2020.0110369.

[17] Liu, Y., Zhang, M., Fan, Z., & Chen, Y. (2020). Heart disease prediction based on random forest and LSTM. *In 2020 2nd international conference on information technology and computer application (ITCA) IEEE*, 630-635.

[18] Sun, L., Wang, Y., He, J., Li, H., Peng, D., & Wang, Y. (2020). A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. *Health information science and systems*, 8, 1-7.

[19] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science*, 165, 74-82.

[20] Sahar, A., & Han, D. (2018). An LSTM-based indoor positioning method using Wi-Fi signals. *In Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, 1-5.

[21] Wei, C. C. (2020). Development of Stacked Long Short-Term Memory Neural Networks with Numerical Solutions for Wind Velocity Predictions. *Adv. Meteorol.*, from doi: 10.1155/2020/5462040.

[22] UCI Machine Learning Repository. (2021). *Heart Disease Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease.

[23] Ridzuan, F., & Zainon, W. M. N. W. (2019). A review of data cleansing methods for big data. *Procedia Computer Science*, 161, 731-738.

[24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 (56): 1929–1958.

[25] Liu, P., Sun, X., Han, Y., He, Z., Zhang, W., & Wu, C. (2022). Arrhythmia classification of LSTM autoencoder based on time series anomaly detection. *Biomedical Signal Processing and Control*, 71, 103228.

[26] Liu, F., Cai, M., Wang, L., & Lu, Y. (2019). An ensemble model based on adaptive noise reducer and over-fitting prevention LSTM for multivariate time series forecasting. *IEEE Access*, 7, 26102-26115.

[27] Hansun, S., Charles, V., & Gherman, T. (2023). The role of the mass vaccination program in combating the COVID-19 pandemic: An LSTM-based analysis of COVID-19 confirmed cases. *Heliyon*, 9(3).

[28] Hansun, S., & Young, J. C. (2021). Predicting LQ45 financial sector indices using RNN-LSTM. *Journal of Big Data*, 8(1), 1-13.

[29] Yoon, T., & Kang, D. (2023). Multi-Modal Stacking Ensemble for the Diagnosis of Cardiovascular Diseases. *Journal of Personalized Medicine*, 13(2).