

## Prediction of patient survival from heart failure using a cox-based model

Tsehay Admassu Assegie<sup>1</sup>, Thulasi Karpagam<sup>2</sup>, Sathya Subramanian<sup>3</sup>,  
Senthil Murugan Janakiraman<sup>4</sup>, Jayanthi Arumugam<sup>5</sup>, Dawed Omer Ahmed<sup>1</sup>

<sup>1</sup>Department of Computer Science, Injibara University, Injibara, Ethiopia

<sup>2</sup>Department of Artificial Intelligence and Data Science, RMK College of Engineering and Technology, Kavaraipettai, India

<sup>3</sup>Department of Electronics and Communication Engineering, Gojan School of Business and Technology, Chennai, India

<sup>4</sup>Department of Computer Science and Engineering, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, India

<sup>5</sup>Department of Computer Science and Engineering, Velammal Engineering College (Autonomous Institution), Affiliated to Anna University, Chennai, India

### Article Info

#### Article history:

Received Jan 25, 2022

Revised Jun 14, 2022

Accepted Jul 3, 2022

#### Keywords:

Cox-model

Electrocardiogram

Heart failure prediction

Survival analysis

Survival prediction

### ABSTRACT

The existing heart failure risk prediction models are developed based on machine learning predictors. The objective of this study is to identify the key risk factors that affect the survival time of heart patients and to develop a heart failure survival prediction model using the identified risk factors. A cox proportional hazard regression method is applied to generate the proposed heart failure survival model. To develop the model multiple risk factors such as age, anemia, creatinine phosphokinase, diabetes history, ejection fraction, presence of high blood pressure, platelet count, serum creatinine, sex, and smoking history. Among the risk factors, high blood pressure is identified as one of the novel risk factors for heart failure. We have validated the performance of the model via statistical and empirical validation. The experimental result shows that the proposed model achieved good discrimination and calibration ability with a C-index (receiver operating characteristic (ROC) of being 0.74 and a log-likelihood ratio of 81.95 using 11 degrees of freedom on the validation dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Tsehay Admassu Assegie

Department of Computer Science, Injibara University

Injibara, Ethiopia

Email: tsehayadmassu2006@gmail.com

## 1. INTRODUCTION

Heart failure is a clinical syndrome characterized by a reduced ability of the heart to pump blood to other parts of the body or fill with blood [1], [2]. Heart failure leads to fatigue, shortness of breath, and poor quality of life. Patients with heart failure have a higher mortality rate, and various bio-statistical methods, as well as machine learning methods, have been applied to predict heart failure deaths from patients' medical records. Savasci *et al.* [3], Toulmi *et al.* [4], and Nugroho *et al.* [5] have conducted on the problem of heart failure diagnosis using different machine learning predictive models. The literature shows that the performance of such a predictive model is acceptable although improving the predictive performance of existing work remained an open research problem. For example, in Su *et al.* [6], the researchers conducted research on improving the performance of the existing heart failure prediction model using hyper-parameter tuning.

In hyper-parameter tuning, the researchers employed a grid search method for determining better parameter settings for effective heart failure prediction using a machine-learning model. The experimental

result shows that the accuracy of the K-nearest neighbor (k-NN) improves by 8.25% with hyperparameter tuning when the optimal value of k is used for training the model. Overall, an accuracy of 86.46% is achieved using the developed model. Literature survey Sumiati *et al.* [7], Animesh *et al.* [8], and Ip *et al.* [9] also shows that different machine learning algorithms are applied to heart failure prediction. In recent years, machine learning is widely adopted in survival analysis [10]-[12]. In the survival analysis, machine learning is applied to determine the issues related to time-to-death events due to heart failure.

The objective of this research is to propose a method for heart failure prediction using the cox proportionate model to investigate determinant risk factors of heart failure. This research explores the answers to the following research questions: i) what is the most significant risk factor for heart failure?; ii) how to optimize a machine-learning model for heart failure survival prediction?; iii) which covariate in the heart failure dataset has a significant impact on the survival rate?; iv) What is the performance of the cox proportionate model for heart failure survival analysis? The rest of the paper is organized as shown in: in section 2, related works are discussed, section 3 presents the method, section 4 discusses the result and findings of the study and finally, section 5 concludes the work.

## 2. METHOD

The dataset employed to conduct this study is collected from University of California Irvine (UCI) heart failure clinical records consisting of 299 samples of which 96 samples are caused death events and 203 non-death events. Cox proportional hazard regression analysis [13]-[15] is employed for developing the proposed heart failure risk survival analysis model, which is one of the most widely, used statistical methods for survival analysis. This research aims to develop a heart failure prediction model using multiple parameters to estimate the probability of surviving heart failure in an individual.

### 2.1. Survival function

Let  $T$  denote the future life of an individual aged 0.  $T$  is a continuous random variable, which takes values on  $\mathbb{R}^+ = [0; \infty)$ . For human life calculation, the limiting age is typically 120 years, thus  $T$  falls in the range  $[0; 120]$ . Then the cumulative distribution function of  $T$  [16]-[19] is given by the probability of death by age  $t$  and is defined by the formula given in (1).

$$F(t) = P(T \leq t) \quad (1)$$

The survival function of  $T$  is the probability of surviving beyond age  $t$  and is defined by (2).

$$S(t) = P(T > t) = 1 - F(t) \quad (2)$$

### 2.2. Exploratory data analysis (EDA)

Exploratory data analysis is recently becoming one of the methods for exporting insight into heart disease datasets [20]-[25]. To explore and get insight into the dataset through understanding the effect of each feature on heart failure survival, the authors employed a seaborn package with a distribution plot. The effect of each of the risk factors in the heart failure dataset is explored for understanding the relationship between each feature and the target variable or the death event due to heart failure. The distribution of numerical covariates and their relationship to death events is demonstrated in Figures 1-5. In Figure 1, the distribution of age and the relationship between age and death events due to heart failure is demonstrated. Most of the death events occurred at the age of 60 as observed in Figure 1.

Figure 1 demonstrates the distribution of age features and the relationship between age features and survival. We observe from Figure 1 that the highest death event occurred at the age of 60 years. In Figure 2, the effect of platelets on heart failure is demonstrated. Figure 3 demonstrates the distribution of creatinine phosphokinase and the relationship between creatinine phosphokinase features and survival. We observe from Figure 3 that the highest death event occurred at the age of 60 years. In Figure 2, the effect of platelets on heart failure is demonstrated.

Figure 3 demonstrates the distribution of creatinine phosphokinase and the relationship between creatinine phosphokinase feature and survival. We observe from Figure 3 that the highest death event occurred at the age of 60 years. In Figure 2, the effect of platelets on heart failure is demonstrated. In Figure 3, the distribution of creatinine phosphokinase and the relationship with the survival of heart disease patients is demonstrated. As demonstrated in Figure 3, lower creatinine phosphokinase is associated with higher death.

In Figure 4, the relationship between the distributions of serum creatinine and the survival of heart disease patients is demonstrated. As demonstrated in Figure 4, lower serum creatinine is associated with higher

death. In Figure 5, the relationship between the distributions of serum sodium and the survival of heart disease patients is demonstrated. As demonstrated in Figure 5, the higher serum sodium is associated with higher death.

In addition to the graphical exploration of the numerical heart failure dataset feature (demonstrated in Figures 1-5), statistical summaries such as count, mean, median, standard deviation max, and min. The dataset contains the following potential risk factors: age, serum sodium, serum creatinine, gender, smoking, blood pressure (BP), ejection fraction (EF), anemia, platelets, and creatinine phosphokinase (CPK). The statistical summary of the dataset is demonstrated in Table 1.

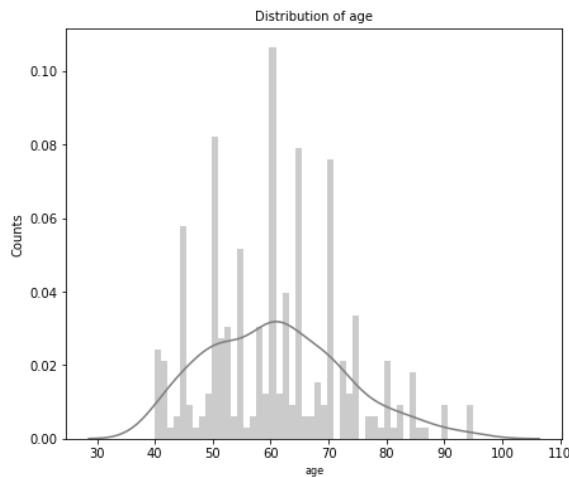


Figure 1. Age distribution and its relation with the heart patient survival

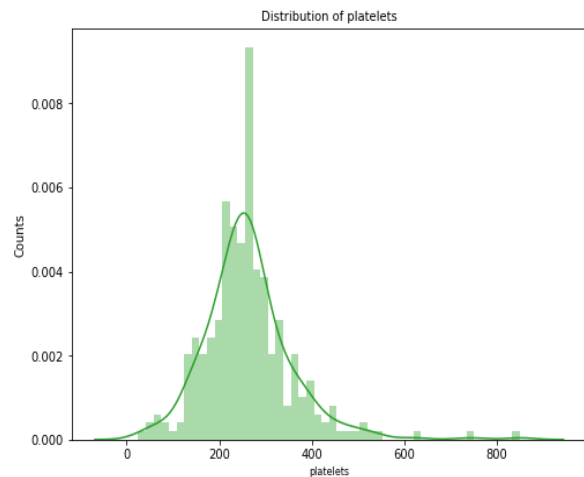


Figure 2. Age distribution and its relation with the heart patient survival

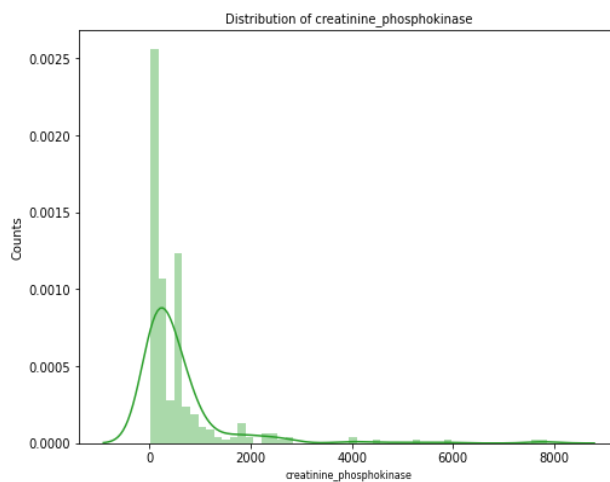


Figure 3. Creatinine phosphokinase and its relation with death event due to heart failure

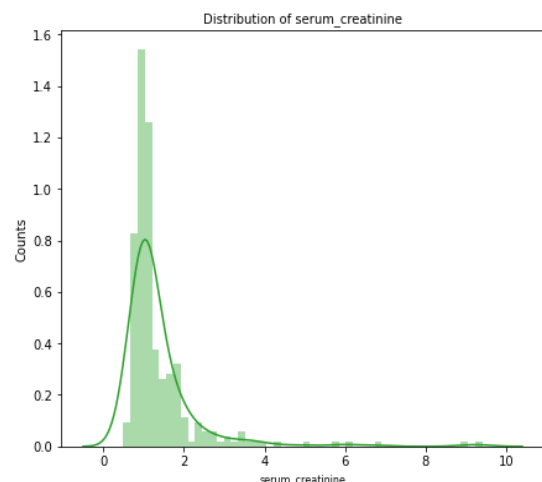


Figure 4. Serum creatinine distribution and its relation with death event due to heart failure

The cox proportional hazard model is fitted on the data respecting censoring, so the model is created by specifying the event to predict i.e. death event due to heart failure, and the time for the occurrence or non-occurrence of the event is observed. In fitting the model, 203 individuals had not died at the time of their follow-up examination at the hospital. From the summary statistics demonstrated in Table 1, we observe that the mean follow-up period was 130 days, ranging from 4 to 285 days.

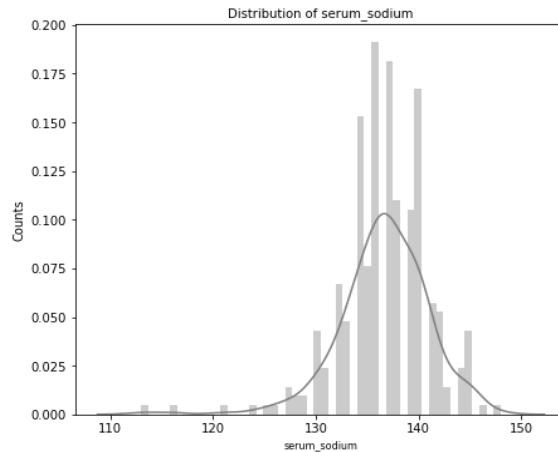


Figure 5. Serum sodium distribution and its relation with death event due to heart failure

Table 1. The performance of the proposed model on heart failure survival analysis

Model	Cox proportionate hazard filter
Duration	Time in days (4 to 285 days)
Event	Death event due to heart failure
baseline estimation	bestow
number of observations	299
Number of events observed	96
partial log-likelihood ratio test	81.95 on 11 degrees of freedom
Concordance	0.741

### 3. RESULTS AND DISCUSSION

The cox model is implemented using the lifelines Python package. After creating the heart failure survival analysis model, we evaluated the model against C-index. To evaluate the performance measure the performance of the proposed model, the Concordance index or C-index is employed. The C-index for the proposed heart failure survival model is the weighted average of the area under time-specific receiver operating characteristic (ROC) curves, which is demonstrated in Table 1. As shown in Table 2, the concordance of the C-index for the proposed model is 0.74, which is a promising result. Hence, the model is effective in predicting the survival time of a patient suffering from heart failure. More statistical results of the proposed model such as p and z values are demonstrated in Table 2.

Table 2. The logrank test performed on covariate variable

Feature	Coef	exp (Coef)	exp (coef lower 95)	exp (coef upper 95)	Z-value	P-value
Age	0.046	1.048	1.029	1.067	4.977	<0.0005
Anaemia	0.460	1.584	1.036	2.423	2.122	0.034
creatinine_phospholipase	0.000	1.000	1.000	1.000	2.226	0.026
diabetes (Nominal)	0.140	1.150	0.743	1.781	0.627	0.531
Bejection_fraction	-0.049	0.952	0.933	0.972	-4.672	<0.0005
high_blood_pressure	0.476	1.609	1.053	2.458	2.201	0.028
platelets	-0.000	1.000	1.000	1.000	-0.412	0.681
Serum creatinine	0.321	1.379	1.201	1.582	4.575	<0.0005
Serum sodium	-0.044	0.957	0.914	1.001	-1.899	0.058
Sex	-0.238	0.789	0.482	1.291	-0.944	0.345
Smoking	0.129	1.138	0.695	1.861	0.513	0.608

As demonstrated in Table 2, age is the most important variable evidenced by the highest z-test. An increase in age will give a 1.029-fold increase in heart failure death event risk. Serum creatinine is the second most important variable evidenced by the 4.575 z-test value. This means, that for each unit increase in serum creatinine the death risk from heart failure is increased by 1.582-fold. Additionally, in figure 6, the risk factors for heart failure death events are demonstrated. We see from Figure 6 that high blood pressure, anemia, serum creatinine, age and diabetes, smoking, and age are all within the 95% confidence interval of affecting death event due to heart failure. We observe from Figure 6 that features such as sex, serum sodium, platelets, ejection

fraction and creatinine phosphokinase are all below the 95% confidence interval of affecting death events due to heart failure.

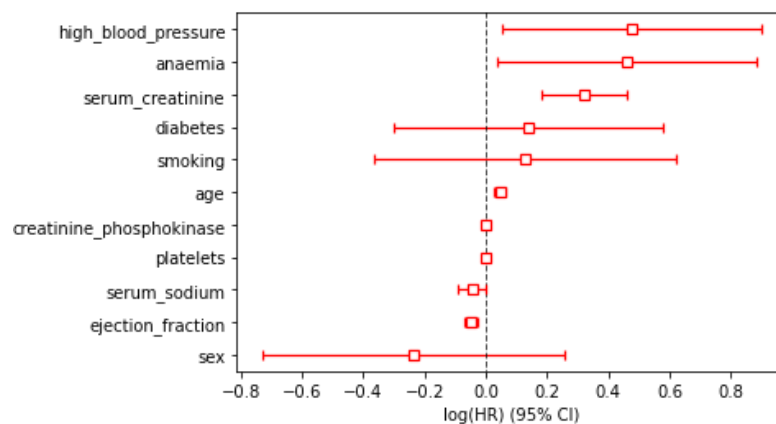


Figure 6. The logarithmic hazard rate (HR) of risk factors of death event due to heart failure

#### 4. CONCLUSION

In this study, a heart failure risk prediction model is proposed using cox proportionate hazard. The proposed model is evaluated against concordance or c-index and the experimental result appears to prove that the model is effective to estimate the survival functions with reasonable accuracy (c-index) for individuals with heart failure. Overall, the proposed model will allow us to estimate how likely a person is to survive or die over time due to heart failure. The proposed model achieved a c-index of 0.74 on the experimental test, which is a promising result.





#### REFERENCES

- [1] X. Jia, M. M. Baig, F. Mirza, and H. GholamHosseini, "A Cox-based risk prediction model for early detection of cardiovascular disease: identification of key risk factors for the development of a 10-year CVD risk prediction," *Advances in Preventive Medicine*, vol. 2019, pp. 1–11, Apr. 2019, doi: 10.1155/2019/8392348.
- [2] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.
- [3] D. Savasci, M. Ceylan, A. H. Ornek, M. Konak, and H. Soylu, "Heart disease detection from neonatal infrared thermograms using multiresolution features and data augmentation," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 8, no. 1, pp. 28–36, Mar. 2020, doi: 10.18201/ijisae.2020158886.
- [4] Y. Toulmi, T. B. Drissi, and B. Nsiri, "Electrocardiogram signals classification using discrete wavelet transform and support vector machine classifier," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 960–970, Dec. 2021, doi: 10.11591/ijai.v10.i4.pp960-970.
- [5] K. S. Nugroho, A. Y. Sukmadewa, A. Vidiyanto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 345–355, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp345-355.
- [6] X. Su *et al.*, "Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model," *Journal of Clinical Laboratory Analysis*, vol. 34, no. 9, Sep. 2020, doi: 10.1002/jcla.23421.
- [7] Sumiati, H. Saragih, T. K. A. Rahman, and A. Triayudi, "Expert system for heart disease based on electrocardiogram data using certainty factor with multiple rule," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 43–50, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp43-50.
- [8] H. Animesh, K. M. Subrata, G. Amit, M. Arkomita, and A. Mukherje, "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review," *Advances in Computational Sciences and Technology*, vol. 10, no. 7, pp. 2137–2159, 2017, [Online]. Available: <http://www.ripublication.com>.
- [9] E. H. Ip, A. Efendi, G. Molenberghs, and A. G. Bertoni, "Comparison of risks of cardiovascular events in the elderly using standard survival analysis and multiple-events and recurrent-events methods," *BMC Medical Research Methodology*, vol. 15, no. 1, p. 15, Dec. 2015, doi: 10.1186/s12874-015-0004-3.
- [10] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, Dec. 2018, doi: 10.1186/s12874-018-0482-1.
- [11] T. A. Assegie, R. L. Tulasi, V. Elanangai, and N. K. Kumar, "Exploring the performance of feature selection method using breast cancer dataset," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 232–237, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp232-237.





- [12] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1831–1838, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.
- [13] O. Rahman *et al.*, "Internet of things based electrocardiogram monitoring system using machine learning algorithm," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 4, pp. 3739–3751, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3739-3751.
- [14] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 184–190, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp184-190.
- [15] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4217–4225, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.
- [16] T. R. Stella Mary and S. Sebastian, "Predicting heart ailment in patients with varying number of features using data mining techniques," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2675–2681, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2675-2681.
- [17] N. Sureja, B. Chawda, and A. Vasant, "A novel salp swarm clustering algorithm for prediction of the heart diseases," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 265–272, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp265-272.
- [18] R. R. K. Al-Taie, B. J. Saleh, A. Y. F. Saedi, and L. A. Salman, "Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5229–5239, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.
- [19] S. Krishnan, P. Magalingam, and R. Ibrahim, "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5467–5476, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.
- [20] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, "An improved feature selection approach for chronic heart disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3501–3506, Dec. 2021, doi: 10.11591/eei.v10i6.3001.
- [21] A. A. Ali, H. S. Hassan, and E. M. Anwar, "Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique," in *2020 12th International Conference on Electrical Engineering, ICEENG 2020*, Jul. 2020, pp. 145–150, doi: 10.1109/ICEENG45378.2020.9171739.
- [22] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.
- [23] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2020, doi: 10.1016/j.jksuci.2020.10.013.
- [24] S. Asadi, S. E. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, p. 103690, Mar. 2021, doi: 10.1016/j.jbi.2021.103690.
- [25] A. Kondababu, V. Siddhartha, B. B. Kumar, and B. Penumutchi, "A comparative study on machine learning based heart disease prediction," *Materials Today: Proceedings*, Feb. 2021, doi: 10.1016/j.matpr.2021.01.475.

## BIOGRAPHIES OF AUTHORS






**Tsehay Admassu Assegie**     obtained his Master's degree in Computer Science from Andhra University Faculty of Science, India 2016. He received B.Sc. from Dilla University, Ethiopia in 2013. His research interest includes machine learning, data mining, bioinformatics, network security, and software-defined networking. He has published over 34 journal articles in international journals and conferences. He can be contacted at email: tsehayadmassu2006@gmail.com.






**Thulasi Karpagam**     is currently working as an assistant professor in the Department of Artificial Intelligence and Data Science at R.M.K College of Engineering and Technology, Kavaraipettai, Chennai, India. Her research areas include cloud computing, machine learning, and big data analytics. She can be contacted at email: karpagamdv83@gmail.com.






**Dr. Sathya Subramanian**    is currently working as an Associate Professor in the Department of Electronics & Communication Engineering at Gojan School of Business and Technology, Chennai, Tamil Nadu, India. Her research interest includes MEMS, Image Processing, Embedded System, and IoT. She can be contacted at email: sathyas7979@gmail.com.






**Dr. Senthil Murugan Janakiraman**    is currently working as Associate Professor in the Department of Computer Science and Engineering at Veltech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi, Chennai-600062, Tamil Nadu, India. His research interest includes Software Engineering, Image Processing. He can be contacted at email: jsenthilmuruganmtech@gmail.com.



**Jayanthi Arumugam**    is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Velammal Engineering College, Surapet, Chennai. Her research interests include Data Mining and Machine Learning. She can be contacted at email: jayanthiarumugamk@gmail.com.



**Dawed Omer Ahmed**    is a Lecturer at the Department of Computer Science. He received his B.Sc., in Computer science from Hawassa University in 2013, and M.Sc., Bahir Dar University in 2017. He is Certified in Advanced Certificate in ICT in Education and Training at the National Institute of Technical Teachers Training and Research, Chennai, Ministry of Human Resource Development, Government of India, Taramani, Chennai-600 113, India in 2018. He can be contacted at email: dawed.daveomer.omer@gmail.com.